

NIKLAS LIDSTRÖMER
HUTAN ASHRAFIAN
EDITORS

Artificial Intelligence in Medicine

Artificial Intelligence in Medicine

Niklas Lidströmer • Hutan Ashrafiān
Editors

Artificial Intelligence in Medicine

With 377 Figures and 127 Tables



Editors

Niklas Lidströmer
Department of Women's and
Children's Health
Karolinska Institutet
Stockholm, Sweden

Hutan Ashrafian
Imperial College London
London, UK

ISBN 978-3-030-64572-4 ISBN 978-3-030-64573-1 (eBook)
ISBN 978-3-030-64574-8 (print and electronic bundle)
<https://doi.org/10.1007/978-3-030-64573-1>

© Springer Nature Switzerland AG 2022

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

May this book contribute to freedom of mind, freedom of speech, gender equality, human rights, cosmopolitan tolerance, environmental progress, scientific progress, medical and healthcare advances, and the rise of the individual's control of their health and body.

We dedicate this new standard reference textbook on artificial intelligence in medicine to all of you who deserve a better healthcare.

Foreword to Artificial Intelligence in Medicine

As the practice of medicine has evolved, so has the torrent of data and our inability, as clinicians, to get our arms around it as we care for patients. There is an extraordinary rescue path in sight – based on the use of artificial intelligence (AI) – that will likely be the most significant transformation for healthcare in generations. Indeed, this has the potential to be the Gutenberg moment for medicine.

While AI has been around for decades, it is only in recent years that deep learning algorithms have paved the way for interpreting medical images. Retrospective studies of large datasets of almost every type of medical image have now shown that machines can be trained to provide rapid and accurate “readings” that are often comparable to that of trained expert clinicians. There are some issues with these comparisons, however, since they have initially performed in silico rather than the real world of medical care. And instead of studying clinicians and AI technologies working in synergy, they typically pit algorithms versus physicians. Nevertheless, this is the first and early phase of substantial progress towards a revolution. Where we will lean on machines for help.

The first order for how AI can be applied in medicine is following the Hippocratic principle of first doing no harm by enhancing patient safety and decreasing medical errors. There are at least five million of these each year in the USA. Given the lack of time to assess patients, their data, their scans, and their slides, it will be a boost to get preliminary readouts from appropriately trained algorithmic interpretation. We are already seeing radiology and ophthalmology deep learning algorithms gaining approval by regulatory agencies for such use cases as brain, chest, and abdominal CT scans, mammograms, chest X-rays, and optical coherence tomograms for retinopathy. Eventually no discipline of medicine will be spared from the influence of AI.

For that to happen, better quality appraisal and substantiation will be needed for the medical community to buy into the change. That takes the assessment of these tools through classical and novel clinical trials, which have been scant to date. Until now, the most randomized trials have been performed by gastroenterologists for endoscopy, particularly real-time machine vision during colonoscopy for detection of polyps. We are fortunate that a recent global effort to provide rigorous standards for AI medical research is budding – the CONSORT-AI and SPIRIT-AI guidelines for trials and those forthcoming STARD-AI and QUADAS-AI for AI’s current mainstay of application in

diagnostics, pathology and imaging – these provide the template for researchers and peer-review journals going forward. The better the supporting evidence, the faster appropriate AI tools in medicine will be integrated into routine clinical practice.

There's the pressing order to alleviate burnout. That was already a global crisis in healthcare resource before the SARS-CoV-2 (COVID-19) pandemic set it. Not only are medical errors potentiated by burnout, but the mental health of clinicians is imperiled, as we have seen with unprecedented levels of clinical depression and suicides. A principal explanation for burnout is the inability to provide care for patients, overwhelmed with keyboard and screen data functions, with limited time to deal with complicated issues. This, too, can be addressed by AI. Using natural language processing and synthetic notes from patient–doctor conversations during a clinic visit are so encouraging that we could be seeing the first phase of keyboard liberation in the next few years.

What is desperately needed is the “gift of time.” The time for the human–human connection between patients and their clinician. This relationship has suffered greatly over recent decades, and we will have the opportunity to get it back, make it stronger than ever. The patient–doctor relationship is the essence of medicine. It is about trust, presence, empathy, and communication, that when you're sick, your doctor has your back. This will never be replaced by AI. It's what I labeled “Deep Medicine” and what should be considered the overarching goal of AI in the years ahead.

That gift of time can be realized not just through interpreting images and keyboard liberation, and synthesizing all of a patient's data from multiple sources, but also by giving patients more autonomy. That is what is starting now, the use of deep learning algorithms to provide screening diagnoses for common medical problems like skin rashes, urinary tract infections, ear infections in children, and even the diagnosis of heart arrhythmias via a smartwatch. That's democratizing medicine. No doubt this trend of promoting some autonomy for patients will further give clinicians time to concentrate on the most complex and serious matters. For all important diagnoses, and before treatment is implemented, the “human-in-the-loop” will be necessary for oversight, never allowing the AI to be the sole autonomous factor. All software is subject to malfunctions and adversarial attacks, which we must be mindful can occur in medicine.

How can all the vast information for AI in medicine be culled together in one resource? I would not have thought it was possible, but Niklas Lidströmer and Hutan Ashrafian have proved me wrong. They have brought together over 300 expert authors and 130 chapters to comprehensively cover the field, with an eminently logical framework of three parts as outlined in their preface. They deserve tremendous credit for this herculean undertaking, as do all of the authors for contributing to this endeavor. While some might question the value of such a textbook in the digital era with such dynamic changes, a book that I have relied upon to learn the nuances of AI, *Deep Learning*, by Ian Goodfellow, Yoshua Bengio, and colleagues, is an exemplar in this field that has been invaluable. I am confident that *Artificial Intelligence in Medicine* will follow suit for all of us in the medical community, as we all must aspire to learn about AI, to understand its nuances, strengths, and limitations.

Hopefully, AI will soon become an essential part of the medical school curriculum and this book will be a core reference.

I am honored to write this foreword and hope this extraordinary resource will help you as we go fast forward with integrating AI tools in our practice and enabling the real care of patients, as Francis Peabody wrote in 1927 in JAMA, “The secret of the care of the patient is in caring for the patient.” Almost 100 years later, a suitable use of AI in clinical practice may allow healthcare to ascend to a higher level of quality and value for patients.

La Jolla, California, USA

Eric J. Topol, MD
Scripps Research

Quotation

The saddest aspect of life right now is that science gathers knowledge faster than society gathers wisdom.

Professor **Isaac Asimov** (Айзек Азимов), 1920–1992

Preface

This is in all essence a book on the future of medicine. Artificial intelligence in medicine (AIM) is without doubt the hottest and most intriguing branch of medicine at the moment. It dwells into all areas and aspects of medicine. Creating human hubris and hypes, and provoking raging debates, fears, and intense hopes, it stretches the limits of medical science and researchers' minds, and raises questions of equality, humanism, emotional intelligence, and our very existence.

Without the ardor of dedicated experts, AIM is a vast territory, hard to vanquish. AIM will bestow medicine with unrivaled new tools, and it is the authors' cordial intention to deliver a broad systematic review of its entire history, scientific foundations, present advances, trends, possibilities, and future challenges.

Hitherto medical education has been lacking a profound and unabridged textbook overview for educational purposes. In the last few years, publications have been exponentially increasing within AIM. Soon, it will be very hard, if doable whatsoever, to compile a comprehensive overview within this field in medicine.

This book offers a wide and profound overview of all the latest advancements within the field of artificial intelligence in medicine. The book both explains the basic concepts of AI, proceeds to the applications of AIM, and then analyzes all of the commonest medical specialties and their AIM advancements systematically.

Although AIM will change medicine radically, the amount of AI taught in medical schooling is astonishingly limited. This book aims to bridge the gaps in profound understanding of AI in medicine for medical students, specialist doctors, and other researchers, whose areas will be hugely affected by AI in the very near future.

This AIM textbook brings the reader into the world of applied AI, from the mathematical foundations through the world of algorithms and computer programs to the ultimate benefits for patients and clinicians. It will be shown how AIM learned a lot from general AI for imaging, such as super-resolution, image reconstruction, image matching, inpainting, age estimation, generative adversarial network (GAN) image generation, and other areas.

Equipped with this book, the reader will no longer look at AI in medicine as a phenomenon happening inside an obscure black box – AI in medicine will be unboxed and opened for the imaginations, dissections, and alterations by students, clinicians, researchers, and other interested academics.

The book contains 130 chapters, as of its first edition, and is divided into three parts: the first is on AI in general; the second is on AI in medicine, covering the lessons for all doctors and the common trunk of AI applicable to all walks of medicine; and the third systematically goes through all medical specialties with as much width and depth as possible.

The first part contains 3 chapters (1–3), the second part contains 27 chapters (4–30), and the third part contains 100 chapters (31–130). These chapter numbers are referred to in this preface.

As we write this preface, we realize that our textbook has become the largest, most updated, unabridged, and comprehensive reference textbook on artificial intelligence in medicine, in the whole scientific world. To reach this altitude has been the outspoken ambitions of our publisher Springer Nature and of ourselves – to give this new and exponentially growing scientific community, for AI in all of healthcare, a standard reference cornerstone work.

The first part of this book contains general aspects of artificial intelligence, that is, the basic concepts primed for clinicians (► [Chap. 1](#)) for further understanding of this book. The pensum herein will likely be everyday ingredients in the self-evident curriculum of future medical students. Besides this chapter, we also present a chapter on what AI can learn from medicine (► [Chap. 2](#)). We also present the mathematical foundations of AI in medicine (► [Chap. 3](#)).

In the second part, we continue into the common trunk of AI in medicine, but still hold onto general aspects, that is, the ideas, concepts, problems, and branches that all medical specialties have in common. At first, we present again a basic concepts chapter (► [Chap. 4](#)), followed by an introductory chapter containing “lessons for all doctors” (► [Chap. 5](#)), and an analysis spread over several chapters on the importance of AI in medicine in general (► [Chap. 6](#)) and why it is so important right now, why it will transform the future healthcare.

Part two further contains chapters that drill into medical science through different angles, which are all common to all specialties, such as medical decision making (► [Chap. 11](#)), medical diagnostics (► [Chap. 12](#)), medical education (► [Chap. 22](#)), medical informatics (► [Chap. 16](#)), medical innovation (► [Chap. 7](#)), security associated to this technology (► [Chap. 20](#)) and AI in medical history (► [Chap. 13](#)), ethical challenges (► [Chap. 9](#)), and philosophical aspects (► [Chap. 26](#)).

In this part, we also present chapters on social, legal, and regulatory aspects (► [Chap. 8](#)), patient safety (► [Chap. 14](#)), contestability (► [Chap. 15](#)), and the patient’s perspective (► [Chap. 24](#)), all dedicated in-depth chapters. Since, of course, data is the gold mine of all AI, the electronic health records (EHRs) (► [Chap. 18](#)) are hence one of the focal points of interest here. A chapter is dedicated to educate on the unsupervised mining of large datasets (► [Chap. 21](#)).

To this common part, we also present chapters on AI in medicine from several other general perspectives, such as personalized medicine and precision medicine (► [Chap. 19](#)), with preserved privacy (► [Chap. 10](#)), and how AI will improve evidence-based medicine (EBM) (► [Chap. 17](#)), and AI in medicine and the evolutionary theory (► [Chap. 23](#)).

The global efforts to provide rigorous standards for AI medical research guidelines are demonstrated in a chapter (► [Chap. 23](#)).

We begin wrapping up the second part of this textbook with a chapter on AIM and meta learning and the concept of “learn to learn” (► [Chap. 29](#)), that is, how AI in healthcare can be taught not only to perform narrow tasks but to also do several tasks, and how it can be used for advanced medical learning, without having to build the model from scratch every time.

The final chapter of the second part is a forward-looking, futuristic chapter on quantum computing in the future of healthcare (► [Chap. 30](#)).

In the third part of the textbook, by far the largest one, we dive down from the high altitudes of medical AI into all medical specialties. These are grouped in the same way as they are normally related to each other clinically, practically, and rationally. Some fields have seen vast advances, others have just begun to scent the first nuances of AI.

The large Mahler-like symphony (like the *Symphony of a Thousand*, his 8th symphony) of the textbook’s third part begins with an advanced introduction, the overture, to the medical specialties and the broadly painted emergence of deep medicine (► [Chap. 31](#)). This first chapter of the third part is followed by chapters on a comprehensive and unabridged range of all medical specialties, systematically arranged.

Some fields have seen massive progress in AI, for example, medical imaging (Chaps. ► [32](#)–► [39](#)), all aspects of radiology (Chaps. ► [32](#)–► [36](#)), dermatological (► [Chap. 39](#)) and ophthalmological (Chaps. ► [110](#) and ► [111](#)) photos, and surgical pathology (► [Chap. 37](#)). Hence, we have drilled into these areas with chapters applying a whole array of angles.

Medical imaging is hence richly represented with a row of chapters ranging from an exposé of explainable methods (► [Chap. 35](#)) in medical imaging and diagnostics, musculoskeletal radiology (► [Chap. 34](#)), and interventional radiology (► [Chap. 32](#)) to the new AI-driven field of transdermal optical imaging (► [Chap. 82](#)); a chapter dedicated to an in-depth analysis of automated deep learning for medical imaging (► [Chap. 33](#)); and specialized chapters such as diagnostic radiology in Covid-19 (► [Chap. 36](#)), acute stroke (► [Chap. 109](#)), and many more.

Dermatology (► [Chap. 39](#)) has extensive representation in the book, and also pathology (Chaps. ► [37](#) and ► [38](#)), both represented directly and indirectly by several chapters. In some areas it has been motivated to dedicate a certain field to an independent chapter, for example, the main chapter on surgical pathology (► [Chap. 37](#)), followed by kidney pathology (► [Chap. 38](#)).

Nephrology (Chaps. ► [40](#) and ► [41](#)) is dealt with in several chapters, such as in the pathology chapter (► [Chap. 38](#)), but also in chapters specializing in hemodialysis (► [Chap. 40](#)), and prediction of kidney function and acute kidney injuries (► [Chap. 41](#)).

The impact of AI in healthcare on world economy is massive, and it will grow immeasurably in the coming decade. Several chapters delve into this from various perspectives and explain the impact on health economics, public health (► [Chap. 42](#)), and business models of healthcare (► [Chap. 43](#)).

This unabridged textbook encompasses both clinical and preclinical medicine. To the latter we can mention molecular biology (► [Chap. 76](#)), stem cell

research and regenerative medicine (► Chap. 79), genomics (► Chap. 77), transplantation biology (► Chap. 79), genomic applications (► Chap. 78), blood transcriptomics (► Chap. 80), and clinical biochemistry (► Chap. 57).

This book has largely been produced under the heavy pressure of the Covid-19 pandemic. This has had several profound impacts. What would have earlier resulted in a vast number of live meetings, and a large carbon dioxide imprint due to flights, has now only resulted in one trip for one of the editors, tour and retour Copenhagen–London, and for this flight it has been CO₂ compensated. It has also impacted gravely on the social lives of all authors, their families, and societies. In many cases, it has also had an effect on the authors' physical health or the health of their families. That this major reference book has been published without any delays is a major and united accomplishment by all contributors.

Several chapters deal with pandemics, and do directly mention and elaborate on Covid-19 from various different angles. This pandemic has been mentioned in the chapters on infectious medicine (► Chap. 95), epidemiology (► Chap. 96), infectious biology (► Chap. 98), radiological (► Chap. 36) and biochemical diagnostics (► Chap. 57), the dynamics of infections (► Chap. 99), immunology (► Chap. 101), and in the various pharmacological chapters as well. There are also chapters for other major infectious diseases, such as malaria (► Chap. 97).

Moreover, this book has good gender balance among its authors, who come from all over the world, and all continents are represented. It is literally the whole world's united and benevolent efforts that are gathered in this great volume. We have a chapter dedicated to gender aspects in AIM in general (► Chap. 28), and also on gender aspects in reproductive health (► Chap. 73).

Since this is a modern textbook, praising a sound gender balance and striving for representation from all continents of the planet, a special chapter deals with AI and its importance for healthcare in Africa (► Chap. 44), and we have also taken environmental factors into consideration, and have dedicated a chapter to AIM and its role in climate change and city pollution (► Chap. 45).

Another insight in this modern approach to a standard reference work is the fact that AI in healthcare is not necessarily developed at only traditional university institutions but at innovative frontline start-ups (► Chap. 7) and also in the hacker cultures – hence in the chapter on AIM and hackathon events (► Chap. 25), this phenomenon is further elaborated.

In the third part of the book, all the chapters have been systematically grouped subjectwise, with subjects related to each other ordered adjacently, for example, the immune system: clinical immunology (► Chap. 101), immunoinformatics (► Chap. 100), immunizations (► Chap. 100), and allergies (► Chap. 102) in one group. This group is followed by the field of hematology (► Chap. 103), which also contains specialized chapters on, for example, anemia (► Chap. 104).

Operating specialties, that is, surgery (► Chap. 61), orthopedics (► Chap. 63), and endoscopy procedures (Chaps. ► 66–69) spanning over several chapters, medical (► Chap. 59) and surgical robotics (► Chap. 60), ear-nose-throat (ENT) medicine (► Chap. 70), maxillofacial surgery (► Chap. 64), and dentistry (► Chap. 65) are placed together. Urology

(► Chap. 62) is also placed in this part, and gynecology and obstetrics (► Chap. 71) are placed close to medical disorders in pregnancy (► Chap. 72) and birth control. These are followed by pediatrics (► Chap. 74) and neonatology (► Chap. 75).

There are two robotics chapters: one introductory (► Chap. 59) and one deeper (► Chap. 60). Likewise, the endoscopy chapters encompass various types, from gastroscopy, with a special chapter on Barrett's esophagus (► Chap. 68), to colorectal polyps in colonoscopy (► Chap. 69), and a chapter on gastroenterology (► Chap. 66), followed by the main chapter on endoscopies (► Chap. 67).

Cancer-related subjects: oncology (► Chap. 90), radiotherapy (► Chap. 91) and cancer care (► Chap. 91), breast cancer (► Chap. 92), brain tumors (► Chap. 123), and cervical cancer (► Chap. 94) have been grouped together and placed in proximity to diagnostic chapters, such as breast thermography (► Chap. 93) and mammography (► Chap. 92).

The book contains a series of pharmaceutical chapters, encompassing subjects ranging from pharmacovigilance (► Chap. 46), 3D modeling, and drug discovery (► Chap. 48) to areas such as clinical trials (► Chap. 47).

The cardiovascular (► Chap. 58), cerebrovascular, and stroke medicine (► Chap. 124) chapters are completed with diagnostic chapters such as the one on acute stroke (► Chap. 109).

Neurology with neurosurgery (► Chap. 119) is also accompanied by chapters on neurodegenerative disorders, Parkinson and Alzheimer (► Chap. 120), amyotrophic lateral sclerosis (ALS) (► Chap. 121), and diagnostic specialties such as clinical neurophysiology and electroencephalography (EEG) (► Chap. 125), and chapters focusing on other neurological conditions such as Ménière's disease (► Chap. 122).

New and future specialties are richly represented with chapters on telemedicine (► Chap. 87), health blogs (► Chap. 81), mHealth, smartphones and apps (► Chap. 88), nanomedicine (► Chap. 84), electronic noses for medical diagnostics (► Chap. 86), wearable and implantable computing (► Chap. 85), longevity medicine (► Chap. 83), and the new advances of clinical biochemistry (► Chap. 57).

The ambition to be unabridged has also led to the inclusion of a chapter on alternative medicine (► Chap. 89).

The psychiatry section contains chapters specializing on eating disorders (► Chap. 118), autism spectrum disorders (► Chap. 113), anxiety and depression (► Chap. 112), schizophrenia (► Chap. 114), post-traumatic stress disorder (PTSD) (► Chap. 117), alcohol and drug dependence (► Chap. 116), but also several therapeutic approaches, such as cognitive behavioral therapy (CBT) (► Chap. 115).

A long range of chapters have emergency medicine as a common denominator, not only the already mentioned stroke and cardiovascular chapters, but also chapters dealing with anesthesiology (► Chap. 105), critical care (► Chap. 106), cardiac arrest (► Chap. 107), and clinical toxicology (► Chap. 108).

The endocrinology (► [Chap. 49](#)) chapter is followed by a chapter dedicated to diabetes (► [Chap. 51](#)) and another focusing on hypertension management (► [Chap. 50](#)).

In ophthalmology (► [Chap. 110](#)), large progress has already been made, and the book contains an extensive chapter in this field and also a second chapter on eye diseases in general practice (► [Chap. 111](#)).

The book's largest chapter is on physiotherapy (► [Chap. 128](#)), and it is also supported by neighboring chapters on rehabilitation medicine (► [Chap. 129](#)) and sports medicine (► [Chap. 130](#)).

The advances in AIM have also boosted the rise of a completely new forensic medicine (Chaps. ► [126](#) and ► [127](#)).

Primary care (► [Chap. 52](#)) is a highly interesting field within AIM, since a long range of specialist competences will be made more accessible here. Both the patients and their primary care doctors and nurses will be given AI-powered tools within a long range of medical specialties. A special chapter on nursing practice, primed for education of nurses (► [Chap. 53](#)), has been included. Patients who see their GPs for, for example, skin conditions will be able to benefit from AI advances in dermatology (► [Chap. 39](#)) much more rapidly. The same goes for many endemic conditions, for example, joint pain, rheumatological disorders (► [Chap. 55](#)), osteoporosis (► [Chap. 56](#)) and respiratory disorders (► [Chap. 54](#)), and many more. The radiological competences will reach out to the primary care setting.

As this book is printed, the preparations for the second edition have already started, and we are looking forward to presenting chapters on subjects such as elderly care and gerontology, sleep medicine, epilepsy, and palliative care and a chapter on the prediction of future pandemics. The pharmaceutical section will be expanded with several chapters, and a new section will engulf several aspects of health economics. Almost 40 newly written chapters are in pipeline.

But now, we proudly present to you the first edition of the world's largest and most updated reference work – we wish to accomplish the urgent need of a *standard* reference textbook – on artificial intelligence in medicine.

We hope this book will show how artificial intelligence in medicine, or *deep medicine* as our foreword writer *Professor Eric Topol* coined it, will boost the awesome power of AI and make medicine better, for all the humans involved. As he writes in the foreword – this is medicine's “Gutenberg moment.”

With this book, as an instrument of knowledge of artificial intelligence in medicine, we endeavor that it can be dually used as a source to crystallize the future, and for a refined tangibility of the present. With this in mind, through our human consciousness and artificially intelligent technologies, we hope to minimize the scourge of pathology to humanity and drive health and well-being across the globe.

Karolinska Institute, Stockholm, Sweden
Imperial College, London, UK
February 2022

Dr. Niklas Lidströmer
Dr. Hutan Ashrafian
The Editors



*Cela est bien dit, répondit Candide, mais il faut cultiver notre jardin.**

*“Excellently observed”, answered Candide; “but let us cultivate our garden”, from *Candide, ou l’Optimisme*, by Voltaire, 1759.

Jean-Jacques Rousseau herborisant à Ermenonville en juin 1778, Moreau, Jean-Michel (dit Moreau le Jeune), Graveur Mayer, Auteur du modèle, ca. 1778. Musée Carnavalet, Histoire de Paris. Published under licence from Musées de la ville de Paris, *La licence Creative Commons Zero (CCØ)*.

Acknowledgments

For invaluable advice and review of the chapters where the editors were involved as authors:

Professor Emeritus Bart ter Haar Romeny

*For the initiated and valuable participation of
All authors of this textbook*

For excellent collaboration with our publisher Springer Nature and the editorial team:

Anusha Cherian

Raasika Dhandapani

Sandra Fabiani

Divya Nithyanandam

Sasikala Rajesh

Aldeena Raju

Melissa Morton

Niels Peter Thomas

Dr. Julia von Graberg

For Dr. Niklas Lidströmer's family:

Louise Lidströmer

Pär Gunnar Thelander

For Dr. Hutan Ashrafian's family:

Dr. Leanne Harling

Ariabella Ashrafian

Persie Ashrafian

For outstanding contribution to nine chapters:

Dr. Joseph D Davids

Claire E Davids

For valuable help with the nursing chapter:

Professor Bertha Ochieng

Gill Meetoo

For valuable input on the physiotherapy chapter:
Jennifer Voller

For engaging Women in AI as its ambassador:
Elena Kell

For enthusiasm and belief in this project:
Professor Yonina Eldar
Professor Eric Herlenius
Professor Eric Topol
Professor Anders Gustafsson
Dr. Richard Dybowski
Dr. Viknesh Sounderajah

For enduring presence and encouragements:
Dr. Tea Kölhi
Lena & Walter Schaller
Felicitas Scholten

For encouragements:
Carola Almqvist, Charlotta Andersson, Anna Asplund with staff, Annette Belfrage, Professor Lennart Dreyer, Simone Fischer, Dr Jessica Hammarström Griffith, Henrik Hedelius, Mikaela Hevring, Johanna Kruse, Paulina Nyquist, Petra Sas, Amy Sevelin, Jan Sjöberg, Johanna Sykora, Patrik Teste, Professor Ola Winqvist

In Memoriam

Our colleague, lecturer, and chapter author **Dr. Danny D Meetoo** joined this reference textbook with academic ardor and vivid enthusiasm. Danny contributed the chapter on AIM and nursing practice, which was created with impressive speed. This chapter was created in joy, together with his wife.

The chapter was delivered in the evening of Saturday, March 6, and a little more than a week later, we received the unbearably sad news that Danny had suddenly and unexpectedly died on Tuesday, March 9.

From 1996, Dr. Danny D Meetoo was a lecturer in the School of Nursing and Midwifery at Salford University. His main specialism was diabetes, with a PhD thesis on the study of diabetes non-compliance and self-care activities. He was prolifically published with articles on diabetes, nanotechnology, and artificial intelligence, and continued to support his former students all over the world, in their research.

At the time of his death, he was proposing a book of his own on the application of AI in nursing. He had started the preliminary work but decided to wait until having moved to a new house before progressing further – he was due to move on March 19.

Dr. Danny D Meetoo was filled with wonder about many things, the beauty of his native island Mauritius, which he left at the age of 17, the wonder of literature, the wonder of how infinite the mind could be for those who chose to open it. He regarded himself as a lifelong student, thereby fulfilling the promise he made to his parents when he left home 54 years ago.

Our thoughts go to his wife and life friend Mrs. Gill Meetoo and the family. Even though words cannot express the shock and sadness, we would like to add this dedication, to the memory of **Dr. Danny D Meetoo**, who has bestowed us with his academic legacy, wisdom, and kindness.

His thoughts, work, and legacy can be reflected with these words by Socrates:

Wisdom begins in wonder

Διά τὸ θαυμάζειν ἡ σοφία

Plato, *Theaetetus*, 155c-d

Contents

Volume 1

Part I	1
1 Basic Concepts of Artificial Intelligence: Primed for Clinicians	3
Niklas Lidströmer, Federica Aresu, and Hutan Ashrafian	
2 Applying Principles from Medicine Back to Artificial Intelligence	21
Howard Schneider	
3 Mathematical Foundations of AIM	37
Yonina C. Eldar, Yuelong Li, and Jong Chul Ye	
Part II	55
4 Introductory Approaches for Applying Artificial Intelligence in Clinical Medicine	57
Niklas Lidströmer, Federica Aresu, and Hutan Ashrafian	
5 Introduction to Artificial Intelligence in Medicine	75
Bart M. ter Haar Romeny	
6 Importance of AI in Medicine	99
Katarina A. M. Gospic and Greg Passmore	
7 The New Frontiers of AI in Medicine	115
Pritesh Mistry	
8 Social and Legal Considerations for Artificial Intelligence in Medicine	129
Matjaž Perc and Janja Hojnik	
9 Ethical Challenges of Integrating AI into Healthcare	139
Lisa Soleymani Lehmann	

10 Artificial Intelligence in Medicine and Privacy Preservation	145
Alexander Ziller, Jonathan Passerat-Palmbach, Andrew Trask, Rickmer Braren, Daniel Rueckert, and Georgios Kaassis	
11 Artificial Intelligence for Medical Decisions	159
Albert Buchard and Jonathan G. Richens	
12 Artificial Intelligence for Medical Diagnosis	181
Jonathan G. Richens and Albert Buchard	
13 AIM and the History of Medicine	203
Kadircan H. Keskinbora	
14 AIM and Patient Safety	215
M. Abdulhadi Alagha, Anastasia Young-Gough, Mataroria Lyndon, Xaviour Walker, Justin Cobb, Leo Anthony Celi, and Debra L. Waters	
15 Right to Contest AI Diagnostics	227
Thomas Ploug and Søren Holm	
16 AIM in Medical Informatics	239
Pierangela Bruno, Francesco Calimeri, and Gianluigi Greco	
17 Artificial Intelligence in Evidence-Based Medicine	255
Artur J. Nowak	
18 AIM in Electronic Health Records (EHRs)	267
Yi Guan and Jingchi Jiang	
19 AIM and Causality for Precision and Value-Based Healthcare	287
Hector Zenil	
20 AIM and the Nexus of Security and Technology	293
Kiran Heer Kaur	
21 AIM in Unsupervised Data Mining	303
Luis I. Lopera González, Adrian Derungs, and Oliver Amft	
22 AIM in Medical Education	319
Joseph Davids, Kyle Lam, Amr Nimer, Stamatia Gianarrou, and Hutan Ashrafian	
23 AIM and Evolutionary Theory	341
Jonathan R. Goodman and Nicolai Wohns	
24 AIM and the Patient's Perspective	351
David Taylor	
25 AIM and Hackathon Events	363
Ayomide Owoyemi and Wuraola Oyewusi	
26 AIM, Philosophy, and Ethics	371
Stephen Rainey, Yasemin J. Erden, and Anais Resseguiher	

27	Reporting Standards and Quality Assessment Tools in Artificial Intelligence–Centered Healthcare Research	385
	Viknesh Sounderajah, Pasha Normahani, Ravi Aggarwal, Shruti Jayakumar, Sheraz R. Markar, Hutan Ashrafiān, and Ara Darzi	
28	AIM and Gender Aspects	397
	Didem Stark and Kerstin Ritter	
29	Meta Learning and the AI Learning Process	407
	Samyakh Tukra, Niklas Lidströmer, and Hutan Ashrafiān	
30	Artificial Intelligence in Medicine Using Quantum Computing in the Future of Healthcare	423
	Joseph Davids, Niklas Lidströmer, and Hutan Ashrafiān	
	Part III	447
31	Emergence of Deep Machine Learning in Medicine	449
	Richard Dybowski	
32	AIM in Interventional Radiology	459
	Suvrankar Datta	
33	Automated Deep Learning for Medical Imaging	473
	Ciara O’Byrne, Laxmi Raja, Robbert Struyven, Edward Korot, and Pearse A. Keane	
34	AI in Musculoskeletal Radiology	487
	Stefan Nehrer, Philip Meier, Matthew D. DiFranco, Zsolt Bertalan, and Richard Ljuhar	
35	AIM and Explainable Methods in Medical Imaging and Diagnostics	501
	Syed Muhammad Anwar	
36	Optimizing Radiologic Detection of COVID-19	511
	Z. Gandomkar, P. C. Brennan, and M. E. Suleiman	
37	AIM in Surgical Pathology	521
	Clare McGenity, Alex Wright, and Darren Treanor	
38	Artificial Intelligence in Kidney Pathology	539
	Sato Noriaki, Uchino Eiichiro, and Okuno Yasushi	
39	AIM in Dermatology	551
	Christian Greis	
40	Artificial Intelligence in Predicting Kidney Function and Acute Kidney Injury	561
	Eiichiro Uchino, Noriaki Sato, and Yasushi Okuno	
41	AIM in Hemodialysis	579
	Oscar J. Pellicer-Valero, Carlo Barbieri, Flavio Mari, and José D. Martín-Guerrero	

42 Artificial Intelligence in Public Health	593
Thomas Lefèvre and Sabine Guez	
43 AIM and Business Models of Healthcare	603
Edward Christopher Dee, Ryan Carl Yu, Leo Anthony Celi, and Umbreena Sultana Nehal	
44 AIM for Healthcare in Africa	613
Ayomide Owoyemi, Adenekan Osiyemi, Joshua Owoyemi, and Andy Boyd	
45 Aim in Climate Change and City Pollution	623
Pablo Torres, Beril Sirmacek, Sergio Hoyas, and Ricardo Vinuesa	
46 AIM in Pharmacology and Drug Discovery	635
Hiroaki Iwata, Ryosuke Kojima, and Yasushi Okuno	
47 Clinical Evaluation of AI in Medicine	645
Xiaoxuan Liu, Gagandeep Sachdeva, Hussein Ibrahim, Maria Charalambides, and Alastair K. Denniston	
48 Artificial Intelligence in Medicine: Biochemical 3D Modeling and Drug Discovery	661
Richard Dybowski	
49 AIM in Endocrinology	673
Namki Hong, Yurang Park, Seng Chan You, and Yumie Rhee	
50 Artificial Intelligence and Hypertension Management	689
Hiroshi Koshimizu and Yasushi Okuno	
51 Aim and Diabetes	701
Josep Vehi, Omer Mujahid, and Ivan Contreras	
52 AIM in Primary Healthcare	711
Niklas Lidströmer, Joseph Davids, Harpreet S. Sood, and Hutan Ashrafiyan	
53 AIM in Nursing Practice	743
Danny D. Meetoo and Bertha Ochieng	
54 AIM in Respiratory Disorders	759
Nilakash Das, Marko Topalovic, and Wim Janssens	
55 AIM in Rheumatology	773
Ching-Heng Lin and Chang-Fu Kuo	
56 AIM in Osteoporosis	785
Sokratis Makrogiannis and Keni Zheng	
57 Artificial Intelligence in Laboratory Medicine	803
Davide Brinati, Luca Ronzio, Federico Cabitza, and Giuseppe Banfi	

58	Artificial Intelligence in Medicine (AIM) in Cardiovascular Disorders	813
	Hisaki Makimoto	
59	AIM in Medical Robotics	825
	Sara Moccia and Elena De Momi	
60	AI in Surgical Robotics	835
	Samyakh Tukra, Niklas Lidströmer, Hutan Ashrafian, and Stamatia Gianarrou	
61	Artificial Intelligence in Surgery	855
	Filippo Filicori and Ozanan R. Meireles	
62	Artificial Intelligence in Urology	863
	Kevin Y. Chu and Michael B. Tradewell	
63	Artificial Intelligence in Trauma and Orthopedics	873
	Roshana Mehdian and Matthew Howard	
64	Harnessing Artificial Intelligence in Maxillofacial Surgery	887
	Karishma Rosann Pereira	
65	AIM in Dentistry	905
	Mauricio do Nascimento Gerhardt, Sohaib Shujaat, and Reinhilde Jacobs	
66	Artificial Intelligence in Gastroenterology	919
	Inga Strümke, Steven A. Hicks, Vajira Thambawita, Debesh Jha, Sravanti Parasa, Michael A. Riegler, and Pål Halvorsen	
67	AIM in Endoscopy Procedures	939
	Aldo Marzullo, Sara Moccia, Francesco Calimeri, and Elena De Momi	
68	AIM in Barrett's Esophagus	951
	Joost van der Putten and Fons van der Sommen	
69	Artificial Intelligence for Colorectal Polyps in Colonoscopy	967
	Luisa F. Sánchez-Peralta, J. Blas Pagador, and Francisco M. Sánchez-Margallo	
70	AIM in Otolaryngology and Head and Neck Surgery	983
	Manish M. George and Neil S. Tolley	
71	AIM in Obstetrics and Gynecology	1003
	Shravanti Muthu, Fatima Nabi, and Junaid Nabi	
72	AIM in Medical Disorders in Pregnancy	1007
	Charles L. Bormann and Carol Lynn Curchoe	

Volume 2

- 73 AIM and Gender Aspects in Reproductive Medicine** 1017
Kiran Heer Kaur
- 74 Artificial Intelligence in Pediatrics** 1029
Christopher J. Kelly, Alexander P. Y. Brown, and
James A. Taylor
- 75 AIM in Neonatal and Pediatric Intensive Care** 1047
David Forsberg, Antoine Honoré, Kerstin Jost, Emma Persad,
Karen Coste, Saikat Chatterjee, Susanne Rautiainen, and
Eric Herlenius
- 76 Aging and Alzheimer's Disease** 1057
Ruixue Ai, Xurui Jin, Bowen Tang, Guang Yang,
Zhangming Niu, and Evandro F. Fang
- 77 Aim in Genomics** 1073
Paola Velardi and Lorenzo Madeddu
- 78 AIM in Genomic Basis of Medicine: Applications** 1087
Mayumi Kamada and Yasushi Okuno
- 79 Stem Cell Progression for Transplantation** 1097
Nazneen Pathan, Sharayu Govardhane, and Pravin Shende
- 80 Artificial Intelligence in Blood Transcriptomics** 1109
Stefanie Warnat-Herresthal, Marie Oestreich,
Joachim L. Schultze, and Matthias Becker
- 81 AIM in Health Blogs** 1125
Paola Velardi and Andrea Lenzi
- 82 AIM and Transdermal Optical Imaging** 1143
Andrew Barszczyk, Weihong Zhou, and Kang Lee
- 83 AI in Longevity Medicine** 1157
Dina Radenkovic, Alex Zhavoronkov, and Evelyne Bischof
- 84 AIM in Nanomedicine** 1169
Joseph Davids and Hutan Ashrafian
- 85 AIM in Wearable and Implantable Computing** 1187
Annalisa Baronetto and Oliver Amft
- 86 Machine Learning and Electronic Noses for Medical
Diagnostics** 1203
Wojciech Wojnowski and Kaja Kalinowska
- 87 Artificial Intelligence in Telemedicine** 1219
Jefferson Gomes Fernandes

-
- 88 AIM and mHealth, Smartphones and Apps** 1229
Joseph Davids and Hutan Ashrafian
- 89 AIM in Alternative Medicine** 1247
Zixin Shu, Ting Jia, Haoyu Tian, Dengying Yan, Yuxia Yang, and Xuezhong Zhou
- 90 AIM in Oncology** 1263
Umar Iqbal and Junaid Nabi
- 91 Artificial Intelligence in Radiotherapy and Patient Care** 1275
James Chun Lam Chow
- 92 Deep Learning in Mammography Breast Cancer Detection** 1287
Richa Agarwal, Moi Hoon Yap, Md. Kamrul Hasan, Reyer Zwiggelaar, and Robert Martí
- 93 AIM for Breast Thermography** 1301
Siva Teja Kakiletli and Geetha Manjunath
- 94 AIM and Cervical Cancer** 1317
Lipi B. Mahanta, Elimah Hussain, and Kangkana Bora
- 95 Artificial Intelligence in Infectious Diseases** 1327
Timothy Miles Rawson, Nathan Peiffer-Smadja, and Alison Holmes
- 96 Artificial Intelligence in Epidemiology** 1341
Thomas Lefèvre and Cyrille Delpierre
- 97 Artificial Intelligence and Malaria** 1353
Cécile Nabet, Aniss Acherar, Antoine Huguenin, Xavier Tannier, and Renaud Piarroux
- 98 Artificial Intelligence in Infection Biology** 1369
Artur Yakimovich
- 99 Artificial Intelligence in Medicine: Modeling the Dynamics of Infectious Diseases** 1379
Richard Dybowski
- 100 AI and Immunoinformatics** 1387
Arash Keshavarzi Arshadi and Milad Salem
- 101 Artificial Intelligence in Clinical Immunology** 1397
Aaron Chin and Nicholas L. Rider
- 102 AIM in Allergy** 1411
Lukas Wisgrill, Paulina Werner, Vittorio Fortino, and Nanna Fyhrquist
- 103 AIM in Haematology** 1425
Joseph Davids and Hutan Ashrafian

- 104 Artificial Intelligence in Medicine in Anemia** 1441
Adam E. Gaweda and Michael E. Brier
- 105 AIM in Anesthesiology** 1453
Matthieu Komorowski and Alexandre Joosten
- 106 Artificial Intelligence in Critical Care** 1469
Alfredo Vellido and Vicent Ribas
- 107 Artificial Intelligence in Medicine (AIM) for Cardiac Arrest** 1479
Hisaki Makimoto
- 108 Artificial Intelligence in Clinical Toxicology** 1487
Meetali Sinha, Praveen G., Deepak Kumar Sachan, and Ramakrishnan Parthasarathi
- 109 Artificial Intelligence in Acute Ischemic Stroke** 1503
Freda Werdiger, Andrew Bivard, and Mark Parsons
- 110 Artificial Intelligence and Deep Learning in Ophthalmology** 1519
Zhaoran Wang, Pearse A. Keane, Michael Chiang, Carol Y. Cheung, Tien Yin Wong, and Daniel Shu Wei Ting
- 111 Artificial Intelligence in Ophthalmology** 1553
Leonardo Seidi Shigueoka, Alessandro Adad Jammal, Felipe Andrade Medeiros, and Vital Paulino Costa
- 112 Aim in Depression and Anxiety** 1567
Kevin Hilbert
- 113 Artificial Intelligence for Autism Spectrum Disorders** 1579
Elisa Ferrari
- 114 Artificial Intelligence in Schizophrenia** 1595
Howard Schneider
- 115 The Rise of the Mental Health Chatbot** 1609
Michiel Rauws
- 116 AIM in Alcohol and Drug Dependence** 1619
Roshan Prakash Rane, Andreas Heinz, and Kerstin Ritter
- 117 Artificial Intelligence in Medicine and PTSD** 1629
Victor Trouset and Thomas Lefèvre
- 118 AIM in Eating Disorders** 1643
D. Kopyto, L. Uhlenberg, R. Zhang, V. Stonawski, S. Horndasch, and Oliver Amft
- 119 AIM in Neurology** 1663
Daisy Das and Lipi B. Mahanta

120	AIM in Neurodegenerative Diseases: Parkinson and Alzheimer	1675
	Joseph Davids and Hutan Ashrafiān	
121	AIM in Amyotrophic Lateral Sclerosis	1691
	Meysam Ahangaran and Adriano Chiò	
122	AIM in Ménière's Disease	1705
	Young Sang Cho and Won-Ho Chung	
123	AIM and Brain Tumors	1717
	Jakub Nalepa	
124	Artificial Intelligence in Stroke	1733
	Nishant K. Mishra and David S. Liebeskind	
125	AIM in Clinical Neurophysiology and Electroencephalography (EEG)	1753
	Joseph Davids, Viraj Bharambe, and Hutan Ashrafiān	
126	Artificial Intelligence in Forensic Medicine	1767
	Thomas Lefèvre	
127	AI in Forensic Medicine for the Practicing Doctor	1777
	Laurent Tournois and Thomas Lefèvre	
128	Artificial Intelligence for Physiotherapy and Rehabilitation	1789
	Joseph Davids, Niklas Lidströmer, and Hutan Ashrafiān	
129	AIM in Rehabilitation	1809
	Parastu Rahgozar	
130	AIM in Sports Medicine	1819
	João Gustavo Claudino, Daniel de Oliveira Capanema, and Paulo Roberto Pereira Santiago	
	Index	1825

About the Editors



Dr. Niklas Lidströmer – Karolinska Institute, MD, MSc, specialist physician, postgraduate researcher in AI in medicine, senior advisor in AI and medical investments, former AI entrepreneur and founder of an AI powered medical platform, former head of Medical AI at a variety of med-tech companies, and also previous co-leader of a handful of successful medical startups.

His experience also encompasses widespread global clinical work spanning 20 years within numerous regions across eight countries. After graduating with a master's thesis on global medicine in 2000, he began practicing as a medical doctor in 2002, followed by internship, specialized residencies, and clinical work all over the world, including 1 year circumnavigating as a maritime doctor.

His international work experience, fluency in nearly ten languages, practical familiarization with AI in the medical and pharmaceutical industries, and clinical specialist competence in general medicine have produced a passion for translational and educational aspects of artificial intelligence in medicine.

Dr. Niklas Lidströmer is eager to bestow upon the world this pivotal reference work – the new standard reference, for artificial intelligence in medicine, which has now become the largest and most comprehensive in the scientific community.



Hutan Ashrafiyan, MBBS, MRCS, PhD, MBA, is a clinician-scientist and active surgeon translating novel technologies and therapeutics in healthcare and policy. He has led R&D as chief scientific adviser at the Institute of Global Health Innovation at Imperial College London and as chief medical officer at a FTSE 100 multinational, and is currently chief scientific officer at the global biotech and venture firm Flagship Pioneering in Preemptive Medicine and Health Security and his own start-ups. He has over 20 years of translational clinical, computational physiology, digital and AI trial, and product development experience, including novel COVID vaccines and national tracing apps. He leads the STARD-AI and QUADAS-AI global guideline initiatives for AI diagnostic accuracy. As honorary lecturer at Imperial College London, he runs the collaboration with Imperial College London, NHS Hospitals, and Google on an AI algorithm for breast screening and also with NICE on health technological assessment classifications for AI. He was awarded the Royal College of Surgeons Arris and Gale Lectureship and the Hunterian Prize. He has authored more than 450 publications (including Lancet, Nature, NEJM) and 10 personally authored books ranging from medicine to philosophy and ancient history, also having discovered an ancient lion species and deciphering hidden realities in Renaissance art including those of Leonardo da Vinci. He has several eponymous medical signs named after him and described his own procedure – the Ashrafiyan Thoracotomy. His philosophical work in artificial general intelligence, human rights, and solving the simulation argument is taught at law schools, and he is regularly featured in historical and scientific documentaries. He has co-edited the major reference book *Artificial Intelligence in Medicine* by Springer Nature.

Contributors

Aniss Acherar Sorbonne Université, Inserm, Institut Pierre-Louis d’Epidémiologie et de Santé Publique, IPLESP, AP-HP, Groupe Hospitalier Pitié-Salpêtrière, Service de Parasitologie-Mycologie, Paris, France

Richa Agarwal Computer Vision and Robotics Institute, University of Girona, Girona, Spain

Ravi Aggarwal Department of Surgery & Cancer, Imperial College London, London, UK

Institute of Global Health Innovation, Imperial College London, London, UK

Meysam Ahangaran Computer Engineering – Artificial Intelligence, Iran University of Science and Technology, Tehran, Iran

Mazandaran University of Science and Technology, Babol, Iran

Ruixue Ai Department of Clinical Molecular Biology, University of Oslo, Oslo, Norway

Akershus University Hospital, Lørenskog, Norway

M. Abdulhadi Alagha MSk Lab, Department of Surgery and Cancer, Imperial College London, London, UK

Institute of Global Health Innovation, Department of Surgery and Cancer, Imperial College London, London, UK

Oliver Amft Digital Health, Friedrich-Alexander University Erlangen-Nuremberg, Erlangen, Germany

Syed Muhammad Anwar Department of Software Engineering, University of Engineering and Technology, Taxila, Pakistan

Federica Aresu KTH Royal Institute of Technology, Stockholm, Sweden

Hutan Ashrafiyan Department of Surgery and Cancer, Imperial College London NHS Trust, London, UK

Institute of Global Health Innovation, Imperial College London, London, UK
Hamlyn Centre for Robotics and Artificial Intelligence, Department of Surgery and Cancer, Imperial College London, London, UK

Giuseppe Banfi IRCCS Istituto Ortopedico Galeazzi, Milano, Italy

Università Vita e Salute San Raffaele, Milano, Italy

Carlo Barbieri Fresenius Medical Care, Bad Homburg, Germany

Annalisa Baronetto Digital Health, FAU Erlangen-Nürnberg, Erlangen, Germany

Andrew Barszczyk Health Management Centre, Drum Tower Hospital Affiliated to Nanjing University Medical School, Nanjing, China

Matthias Becker Systems Medicine, Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE), Bonn, Germany

Zsolt Bertalan ImageBiopsy Lab, Research & AI Development Abteilung, Wien, Austria

Viraj Bharambe Walton Centre for Neurology and Neurosurgery, Liverpool, UK

Evelyne Bischof International Center for Multimorbidity and Complexity in Medicine (ICMC), Universität Zürich, Schweiz, Zurich, Switzerland

College of Clinical Medicine, Shanghai University of Medicine and Health Sciences, Shanghai, China

Human Longevity, Inc., San Diego, CA, USA

Andrew Bivard Melbourne Brain Centre at Royal Melbourne Hospital, Melbourne, VIC, Australia

Department of Medicine, University of Melbourne, Melbourne, VIC, Australia

Kangkana Bora Cotton University, Guwahati, India

Charles L. Bormann Massachusetts General Hospital IVF Laboratory, Boston, MA, USA

Andy Boyd University of Illinois at Chicago, Chicago, IL, USA

Rickmer Braren Institute for Diagnostic and Interventional Radiology, School of Medicine, Technical University of Munich, Munich, Germany

P. C. Brennan University of Sydney, Sydney, NSW, Australia

Michael E. Brier Division of Nephrology and Hypertension, University of Louisville, Louisville, KY, USA

Robley Rex Veterans Administration Medical Center, Louisville, KY, USA

Davide Brinati IRCCS Istituto Ortopedico Galeazzi, Milano, Italy

Alexander P. Y. Brown Google Health, London, UK

Pierangela Bruno Department of Mathematics and Computer Science, University of Calabria, Rende, Italy

Albert Buchard Service de Psychiatrie adulte, Hopitaux Universitaires de Genve, Geneva, CH, Switzerland

Federico Cabitza Università Vita e Salute San Raffaele, Milano, Italy

Francesco Calimeri Department of Mathematics and Computer Science, University of Calabria, Rende, Italy

Daniel de Oliveira Capanema Computing Department, Federal Center for Technological Education of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

Leo Anthony Celi Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA

Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, MA, USA

Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA

Maria Charalambides College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK

Saikat Chatterjee Division of Information Science and Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

Carol Y. Cheung Department of Ophthalmology and Visual Sciences, The Chinese University of Hong Kong, Hong Kong, China

Michael Chiang Departments of Ophthalmology and Medical Informatics and Clinical Epidemiology, Casey Eye Institute, Oregon Health and Science University, Portland, OR, USA

Aaron Chin Department of Medicine, Baylor College of Medicine, Houston, TX, USA

Adriano Chiò ‘Rita Levi Montalcini’ Department of Neuroscience, University of Turin, Turin, Italy

Young Sang Cho Department of Otorhinolaryngology-Head and Neck Surgery, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, South Korea

James Chun Lam Chow Radiation Medicine Program, Princess Margaret Cancer Centre, University Health Network, Toronto, ON, Canada

Department of Radiation Oncology, University of Toronto, Toronto, ON, Canada

TECHNA Institute for the Advancement of Technology for Health, University Health Network, Toronto, ON, Canada

Kevin Y. Chu Department of Urology, University of Miami Miller School of Medicine, Miami, FL, USA

Won-Ho Chung Department of Otorhinolaryngology-Head and Neck Surgery, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, South Korea

João Gustavo Claudino Research and Development Department, LOAD CONTROL, Contagem, Minas Gerais, Brazil

School of Physical Education and Sport – Laboratory of Biomechanics, Universidade de São Paulo, São Paulo, São Paulo, Brazil

Justin Cobb MSk Lab, Department of Surgery and Cancer, Imperial College London, London, UK

Ivan Contreras Modelling, Identification and Control Engineering Laboratory (MICELab), Institut d'Informatica i Applicacions, Universitat de Girona, Girona, Spain

Vital Paulino Costa University of Campinas, São Paulo, Brazil

Karen Coste Department of Women's & Children's Health, Karolinska Institutet, Stockholm, Sweden

Astrid Lindgren Children's Hospital, Children's and Women's Health, Karolinska University Hospital, Stockholm, Sweden

CNRS, INSERM, GReD, Université Clermont Auvergne, Clermont-Ferrand, France

Carol Lynn Curchoe Fertility Guidance Technologies, Newport Beach, CA, USA

Ara Darzi Department of Surgery & Cancer, Imperial College London, London, UK

Institute of Global Health Innovation, Imperial College London, London, UK

Daisy Das Institute of Advanced Study in Science and Technology, Guwahati, India

Nilakash Das Laboratory of Respiratory Diseases and Thoracic Surgery, Department of Chronic Diseases, Metabolism and Ageing, Katholieke Universiteit Leuven, Leuven, Belgium

Suvrankar Datta Department of Radiodiagnosis and Interventional Radiology, All India Institute of Medical Sciences (AIIMS), New Delhi, India

Joseph Davids Imperial College London NHS Trust, London, UK

Hamlyn Centre for Robotics and Artificial Intelligence, Department of Surgery and Cancer, Imperial College London, London, UK

National Hospital for Neurology and Neurosurgery Queen Square, London, UK

Elena De Momi Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy

Edward Christopher Dee Harvard Medical School, Boston, MA, USA

Cyrille Delpierre CERPOP, Center for Epidemiology and Research in Population Health, Université de Toulouse, Inserm, UPS, Toulouse, France

Alastair K. Denniston University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK

Centre for Regulatory Science and Innovation, Birmingham Health Partners, University of Birmingham, Birmingham, UK

College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK

Health Data Research UK, London, UK

National Institute of Health Research Biomedical Research Centre for Ophthalmology, Moorfields Hospital London NHS Foundation Trust and University College London, Institute of Ophthalmology, London, UK

Adrian Derungs Friedrich Alexander University Erlangen-Nuremberg, Erlangen, Germany

Matthew D. DiFranco ImageBiopsy Lab, Research & AI Development Abteilung, Wien, Austria

Mauricio do Nascimento Gerhardt OMFS IMPATH Research Group, Department of Imaging and Pathology, University of Leuven and Oral & Maxillofacial Surgery, University Hospitals Leuven, KU Leuven, Leuven, Belgium

School of Health Sciences, Faculty of Dentistry, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, Brazil

Richard Dybowski St John's College, Cambridge, UK

Uchino Eiichiro Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto University, Kyoto, Japan

Department of Nephrology, Graduate School of Medicine, Kyoto University, Kyoto, Japan

Yonina C. Eldar Department of Math and Computer Science, Weizmann Institute of Science, Rehovot, Israel

Yasemin J. Erden University of Twente, Enschede, The Netherlands

Evandro F. Fang Department of Clinical Molecular Biology, University of Oslo and Akershus University Hospital, Lørenskog, Norway

The Norwegian Centre on Healthy Ageing (NO-Age), Oslo, Norway

Jefferson Gomes Fernandes Telemedicine Education Program, Paulista Medical Association, São Paulo, Brazil

Brazilian Association of Telemedicine and Telehealth, Rio de Janeiro, Brazil
Education Program, International Society for Telemedicine and eHealth, Geneva, Switzerland

CEO, SPECIS – a Health Consulting Firm, São Paulo, Brazil

Elisa Ferrari Scuola Normale Superiore, Pisa, Italy

Filippo Filicori Intraoperative Performance Analytics Laboratory, Department of Surgery, Lenox Hill Hospital, Hofstra School of Medicine at Northwell, New York, NY, USA

David Forsberg Department of Women's & Children's Health, Karolinska Institutet, Stockholm, Sweden

Astrid Lindgren Children's Hospital, Children's and Women's Health, Karolinska University Hospital, Stockholm, Sweden

Vittorio Fortino Institute of Biomedicine, University of Eastern Finland, Kuopio, Finland

Nanna Fyhrquist Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden

Praveen G. Computational Toxicology Facility, CSIR- Indian Institute of Toxicology Research, Lucknow, Uttar Pradesh, India

Z. Gandomkar University of Sydney, Sydney, NSW, Australia

Adam E. Gaweda Division of Nephrology and Hypertension, University of Louisville, Louisville, KY, USA

Manish M. George Imperial College NHS Healthcare Trust, London, UK
ENT Department, St. Mary's Hospital, London, UK

Stamatia Gianarrou Hamlyn Centre for Robotic Surgery and AI, Department of Surgery and Cancer, Imperial College London, London, UK

Jonathan R. Goodman Leverhulme Centre for Human Evolutionary Studies, University of Cambridge, Cambridge, UK

Darwin College, University of Cambridge, Cambridge, UK

Katarina A. M. Gospic Brainbow Labs AB, Stockholm, Sweden

Sharayu Govardhane Shobhaben Pratapbhai Patel School of Pharmacy and Technology Management, SVKM'S NMIMS, Vile Parle (West), Mumbai, India

Gianluigi Greco Department of Mathematics and Computer Science, University of Calabria, Rende, Italy

Christian Greis Department of Dermatology, University Hospital Zurich, Zurich, Switzerland

Yi Guan WI (Web Intelligence) Lab, School of Computer Science and Technology, Harbin, China

Sabine Guez IRIS Institut de Recherche Interdisciplinaire sur les enjeux Sociaux, UMR8156 CNRS – U997 Inserm – EHESS, Université Sorbonne Paris Nord, Paris, France

Pål Halvorsen SimulaMet, Oslo, Norway

Department of Computer Science, Oslo Metropolitan University, Oslo, Norway

Md. Kamrul Hasan Bangladesh University of Engineering and Technology, Dhaka, Bangladesh

Kiran Heer Kaur Center for Social Development, Wolverhampton, UK

Andreas Heinz Department of Psychiatry and Psychotherapy, Charité – Universitätsmedizin Berlin, Berlin, Germany

Eric Herlenius Department of Women's & Children's Health, Karolinska Institutet, Stockholm, Sweden

Astrid Lindgren Children's Hospital, Children's and Women's Health, Karolinska University Hospital, Stockholm, Sweden

Steven A. Hicks SimulaMet, Oslo, Norway

Department of Computer Science, Oslo Metropolitan University, Oslo, Norway

Kevin Hilbert Department of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany

Janja Hojnik Faculty of Law, University of Maribor, Maribor, Slovenia

Søren Holm Centre for Social Ethics and Policy, School of Law, University of Manchester, Manchester, UK

Center for Medical Ethics, Faculty of Medicine, University of Oslo, Oslo, Norway

Alison Holmes Centre for Antimicrobial Optimisation, Imperial College London, London, UK

Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, Imperial College London, Hammersmith Hospital, London, UK

Namki Hong Department of Internal Medicine, Endocrine Research Institute, Yonsei University College of Medicine, Seoul, South Korea

Antoine Honoré Department of Women's & Children's Health, Karolinska Institutet, Stockholm, Sweden

Astrid Lindgren Children's Hospital, Children's and Women's Health, Karolinska University Hospital, Stockholm, Sweden

Division of Information Science and Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

S. Horndasch UK Erlangen, Friedrich-Alexander University Erlangen-Nuremberg, Erlangen, Germany

Matthew Howard Musgrove Park Hospital, Taunton, UK

Sergio Hoyas Instituto Universitario de Matemática Pura y Aplicada, Universitat Politècnica de Valencia, Valencia, Spain

Antoine Huguenin EA 7510, ESCAPE, Laboratoire de Parasitologie-Mycologie, Université de Reims Champagne-Ardenne, Reims, France

Elima Hussain Institute of Advanced Study in Science and Technology, Guwahati, India

Hussein Ibrahim University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK

Centre for Regulatory Science and Innovation, Birmingham Health Partners, University of Birmingham, Birmingham, UK

Umar Iqbal ATLAS Program, Department of Urology, Roswell Park Cancer Institute, Buffalo, NY, USA

Hiroaki Iwata Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto University, Kyoto, Japan

Reinhilde Jacobs OMFS IMPATH Research Group, Department of Imaging and Pathology, University of Leuven and Oral & Maxillofacial Surgery, University Hospitals Leuven, KU Leuven, Leuven, Belgium

Department of Dental Medicine, Karolinska Institute, Stockholm, Sweden

Alessandro Adad Jammal University of Campinas, São Paulo, Brazil

Wim Janssens Laboratory of Respiratory Diseases and Thoracic Surgery, Department of Chronic Diseases, Metabolism and Ageing, Katholieke Universiteit Leuven, Leuven, Belgium

Shruti Jayakumar Department of Surgery & Cancer, Imperial College London, London, UK

Debesh Jha SimulaMet, Oslo, Norway

Department of Computer Science, UIT The Arctic University of Norway, Oslo, Norway

Ting Jia Institute of Medical Intelligence, School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China

Jingchi Jiang WI (Web Intelligence) Lab, School of Computer Science and Technology, Harbin, China

Xurui Jin MindRank AI Ltd., Hangzhou, Zhejiang, China

Alexandre Joosten Department of Anesthesiology and Intensive Care, Hôpitaux Universitaires Paris-Sud, Université Paris-Sud, Université Paris-Saclay, Hôpital De Bicêtre, Assistance Publique Hôpitaux de Paris (AP-HP), Le Kremlin-Bicêtre, France

Kerstin Jost Department of Women's & Children's Health, Karolinska Institutet, Stockholm, Sweden

Astrid Lindgren Children's Hospital, Children's and Women's Health, Karolinska University Hospital, Stockholm, Sweden

Georgios Kaassis Institute for Diagnostic and Interventional Radiology, School of Medicine, Technical, University of Munich, Munich, Germany

Institute for Artificial Intelligence in Medicine and Healthcare, School of Medicine and Department of Informatics, Technical University of Munich, Munich, Germany

OpenMined, Oxford, UK

Siva Teja Kakileti Niramai Health Analytix Private Limited, Bangalore, India

Kaja Kalinowska Department of Analytical Chemistry, Gdańsk University of Technology, Gdańsk, Poland

Mayumi Kamada Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto University, Kyoto, Japan

Pearse A. Keane NIHR Biomedical Research Centre for Ophthalmology, Moorfields Eye Hospital NHS Foundation Trust, London, UK

Christopher J. Kelly Google Health, London, UK

Arash Keshavarzi Arshadi University of Central Florida, Orlando, FL, USA

Kadircan H. Keskinbora School of Medicine, Bahcesehir University, Istanbul, Turkey

Ryosuke Kojima Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto University, Kyoto, Japan

Matthieu Komorowski Department of Surgery and Cancer, Imperial College London, London, UK

Intensive Care Unit, Charing Cross Hospital, Imperial College Healthcare NHS Trust, London, UK

D. Kopyto Digital Health, Friedrich-Alexander University Erlangen-Nuremberg, Erlangen, Germany

Edward Korot Moorfields Eye Hospital NHS Foundation Trust, London, UK

Byers Eye Institute, Stanford University, Palo Alto, CA, USA

Hiroshi Koshimizu Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto University, Kyoto, Japan

Development Center, Omron Healthcare Co., Ltd., Kyoto, Japan

Chang-Fu Kuo Center for Artificial Intelligence in Medicine, Chang Gung Memorial Hospital, Taoyuan, Taiwan

Division of Rheumatology, Allergy, and Immunology, Chang Gung Memorial Hospital, Taoyuan, Taiwan

Kyle Lam Hamlyn Centre for Robotic Surgery and Artificial Intelligence, Department of Surgery and Cancer, Imperial College London, London, UK

Imperial College London NHS Trust, London, UK

Kang Lee Dr. Eric Jackman Institute of Child Study, University of Toronto, Toronto, ON, Canada

Thomas Lefèvre IRIS Institut de Recherche Interdisciplinaire sur les enjeux Sociaux, UMR8156 CNRS – U997 Inserm – EHESS – Université Sorbonne Paris Nord, Paris, France

Department of Forensic and Social Medicine, AP-HP, Jean Verdier Hospital, Bondy, France

Lisa Soleymani Lehmann Google, Harvard Medical School, Brigham and Women's Hospital, Boston, MA, USA

Andrea Lenzi Department of Computer Science, Sapienza University of Rome, Rome, Italy

Yuelong Li Amazon, San Jose, CA, USA

Niklas Lidströmer Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden

David S. Liebeskind Vascular Neurology, University of California, Los Angeles, CA, USA

Ching-Heng Lin Center for Artificial Intelligence in Medicine, Chang Gung Memorial Hospital, Taoyuan, Taiwan

Xiaoxuan Liu University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK

Centre for Regulatory Science and Innovation, Birmingham Health Partners, University of Birmingham, Birmingham, UK

College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK

Richard Ljuhar ImageBiopsy Lab, Research & AI Development Abteilung, Wien, Austria

Luis I. Lopera González Friedrich Alexander University Erlangen-Nürnberg, Erlangen, Germany

Mataroria Lyndon Centre for Medical and Health Science Education, University of Auckland, Auckland, New Zealand

Lorenzo Madeddu Dipartimento di Medicina Traslazionale e di Precisione, Sapienza Università di Roma, Rome, Italy

Lipi B. Mahanta Institute of Advanced Study in Science and Technology, Guwahati, India

Hisaki Makimoto Arrhythmia Service, Division of Cardiology, Pulmonology and Vascular Medicine, Faculty of Medicine, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany

Sokratis Makrogianis Division of Physics, Engineering, Mathematics and Computer Science, Delaware State University, Dover, DE, USA

Geetha Manjunath Niramai Health Analytix Private Limited, Bangalore, India

Flavio Mari Fresenius Medical Care, Bad Homburg, Germany

Sheraz R. Markar Department of Surgery & Cancer, Imperial College London, London, UK

Robert Martí Computer Vision and Robotics Institute, University of Girona, Girona, Spain

José D. Martín-Guerrero Intelligent Data Analysis Laboratory, Department of Electronic Engineering, ETSE (Engineering School), Universitat de València (UV), Bujassot, Valencia, Spain

Aldo Marzullo Department of Mathematics and Computer Science, University of Calabria, Rende, Italy

Clare McGinity Leeds Teaching Hospitals NHS Trust, Leeds, UK
University of Leeds, Leeds, UK

Felipe Andrade Medeiros Department of Ophthalmology, Duke Eye Center, Durham, NC, USA

Danny D. Meetoo School of Nursing, Midwifery and Social Work, University of Salford, Greater Manchester, UK

Roshana Mehdian St Georges Hospital London NHS, London, UK

Philip Meier ImageBiopsy Lab, Research & AI Development Abteilung, Wien, Austria

Ozanan R. Meireles Surgical Artificial Intelligence and Innovation Laboratory, Department of Surgery, Massachusetts General Hospital, Boston, MA, USA

Nishant K. Mishra Department of Neurology, UCLA Stroke Center, University of California, Los Angeles, CA, USA

Pritesh Mistry The Kings Fund, London, UK

Sara Moccia The BioRobotics Institute, Scuola Superiore Sant'Anna, Pisa, Italy

Department of Excellence in Robotics & AI, Scuola Superiore Sant'Anna, Pisa, Italy

Omer Mujahid Modelling, Identification and Control Engineering Laboratory (MICELab), Institut d'Informatica i Aplicacions, Universitat de Girona, Girona, Spain

Shravanti Muthu Boston University School of Public Health, Boston, MA, USA

Cécile Nabet Sorbonne Université, Inserm, Institut Pierre-Louis d'Epidémiologie et de Santé Publique, IPLESP, AP-HP, Groupe Hospitalier Pitié-Salpêtrière, Service de Parasitologie-Mycologie, Paris, France

Fatima Nabi Boston, MA, USA

Junaid Nabi Harvard University, Boston, MA, USA

Jakub Nalepa Silesian University of Technology, Gliwice, Poland

Future Processing Healthcare, Gliwice, Poland

Umbereen Sultana Nehal MIT Sloan School of Management, Cambridge, MA, USA

University of Massachusetts Medical School, Worcester, MA, USA

Stefan Nehrer Donau-Universität Krems, Zentrum für Regenerative Medizin, Krems, Austria

Amr Nimer Imperial College London NHS Trust, London, UK

Zhangming Niu MindRank AI Ltd., Hangzhou, Zhejiang, China

Aladdin Healthcare Technologies Ltd., London, UK

Sato Noriaki Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto University, Kyoto, Japan

Department of Nephrology, Graduate School of Medicine, Kyoto University, Kyoto, Japan

Pasha Normahani Department of Surgery & Cancer, Imperial College London, London, UK

Artur J. Nowak Evidence Prime, Krakow, Poland

Ciara O'Byrne Moorfields Eye Hospital NHS Foundation Trust, London, UK

Trinity College Dublin, Dublin, Ireland

Bertha Ochieng Integrated Health and Social Care, Faculty of Health & Life Sciences, De Montfort University, Leicester, UK

Marie Oestreich Systems Medicine, Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE), Bonn, Germany

Yasushi Okuno Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto University, Kyoto, Japan

Adenekan Osiyemi University of Ibadan Teaching Hospital Ibadan, Ibadan, Nigeria

Ayomide Owoyemi University of Illinois at Chicago, Chicago, IL, USA

Joshua Owoyemi Elix Incorporated, Tokyo, Japan

Wuraola Oyewusi Research and Innovation, Data Science Nigeria, Lagos, Nigeria

J. Blas Pagador Bioengineering and Health Technologies, Jesús Usón Minimally Invasive Surgery Centre, Cáceres, Spain

Sravanthi Parasa Department of Gastroenterology, Swedish Medical Group, Seattle, WA, USA

Yurang Park Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, South Korea

Mark Parsons UNSW South Western Sydney Clinical School Department of Neurology, Liverpool Hospital, Ingham Institute for Applied Medical Research, Liverpool, VIC, Australia

Ramakrishnan Parthasarathi Computational Toxicology Facility, CSIR-Indian Institute of Toxicology Research, Lucknow, Uttar Pradesh, India
Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, Uttar Pradesh, India

Jonathan Passerat-Palmbach OpenMined, Oxford, UK
Department of Computing, Imperial College London, London, UK
Consensys Health, New York, NY, USA

Greg Passmore VR Media Technology, Los Angeles, CA, USA

Nazneen Pathan Shobhaben Pratapbhai Patel School of Pharmacy and Technology Management, SVKM'S NMIMS, Vile Parle (West), Mumbai, India

Nathan Peiffer-Smadja Centre for Antimicrobial Optimisation, Imperial College London, London, UK
Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, Imperial College London, Hammersmith Hospital, London, UK
Université de Paris, Inserm, IAME, Paris, France

Oscar J. Pellicer-Valero Intelligent Data Analysis Laboratory, Department of Electronic Engineering, ETSE (Engineering School), Universitat de València (UV), Bujassot, Valencia, Spain

Matjaž Perc Faculty of Natural Sciences and Mathematics, University of Maribor, Maribor, Slovenia
China Medical University Hospital, China Medical University, Taichung, Taiwan
Complexity Science Hub Vienna, Vienna, Austria

Karishma Rosann Pereira KNR University of Health Sciences, Hyderabad, India

Emma Persad Department of Women's & Children's Health, Karolinska Institutet, Stockholm, Sweden
Astrid Lindgren Children's Hospital, Children's and Women's Health, Karolinska University Hospital, Stockholm, Sweden
Karl Landsteiner University of Health Sciences, Krems, Austria

Renaud Piarroux Sorbonne Université, Inserm, Institut Pierre-Louis d'Epidémiologie et de Santé Publique, IPLESP, AP-HP, Groupe Hospitalier Pitié-Salpêtrière, Service de Parasitologie-Mycologie, Paris, France

Thomas Ploug Centre for Applied Ethics and Philosophy of Science, Department of Communication, Aalborg University Copenhagen, Copenhagen, Denmark

Dina Radenkovic Hooke London, London, UK

King's College London, London, UK

Buck Institute for Research on Aging, Novato, CA, USA

Parastu Rahgozar KTH Royal Institute of Technology, Stockholm, Sweden

Stephen Rainey University of Oxford, Oxford, UK

Laxmi Raja Moorfields Eye Hospital NHS Foundation Trust, London, UK

University College London, London, UK

Roshan Prakash Rane Department of Psychiatry and Psychotherapy, Charité – Universitätsmedizin Berlin, Berlin, Germany

Susanne Rautiainen Department of Women's & Children's Health, Karolinska Institutet, Stockholm, Sweden

Astrid Lindgren Children's Hospital, Children's and Women's Health, Karolinska University Hospital, Stockholm, Sweden

Michiel Rauws X2 AI, San Francisco, CA, USA

Timothy Miles Rawson Centre for Antimicrobial Optimisation, Imperial College London, London, UK

Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, Imperial College London, Hammersmith Hospital, London, UK

Anais Resseguier Trilateral Research, Waterford, Ireland

Yumie Rhee Department of Internal Medicine, Endocrine Research Institute, Yonsei University College of Medicine, Seoul, South Korea

Vicent Ribas Eurecat, Barcelona, Spain

Jonathan G. Richens AI Research, Babylon Health, London, UK

Nicholas L. Rider Section of Immunology, Allergy & Retrovirology, Texas Children's Hospital, Baylor College of Medicine, Houston, TX, USA

Michael A. Riegler SimulaMet, Oslo, Norway

Kerstin Ritter Department of Psychiatry and Psychotherapy, Charité – Universitätsmedizin Berlin, Berlin, Germany

Humboldt-Universität zu Berlin, Berlin, Germany

Department of Psychiatry and Psychotherapy, Berlin Institute of Health, Bernstein Center for Computational Neuroscience, Berlin, Germany

Luca Ronzio Università Vita e Salute San Raffaele, Milano, Italy

Daniel Rueckert Institute for Artificial Intelligence in Medicine and Healthcare, School of Medicine and Department of Informatics, Technical University of Munich, Munich, Germany

Department of Computing, Imperial College London, London, UK

Deepak Kumar Sachan Computational Toxicology Facility, CSIR- Indian Institute of Toxicology Research, Lucknow, Uttar Pradesh, India

Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, Uttar Pradesh, India

Gagandeep Sachdeva College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK

Milad Salem University of Central Florida, Orlando, FL, USA

Francisco M. Sánchez-Margallo Scientific Direction, Jesús Usón Minimally Invasive Surgery Centre, Cáceres, Spain

Luisa F. Sánchez-Peralta Bioengineering and Health Technologies, Jesús Usón Minimally Invasive Surgery Centre, Cáceres, Spain

Paulo Roberto Pereira Santiago School of Physical Education and Sport of Ribeirão Preto – LaBioCoM Biomechanics and Motor Control Laboratory, Universidade de São Paulo, Ribeirão Preto, São Paulo, Brazil

Noriaki Sato Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto University, Kyoto, Japan

Department of Nephrology, Graduate School of Medicine, Kyoto University, Kyoto, Japan

Howard Schneider Sheppard Clinic North, Toronto, ON, Canada

Joachim L. Schultze Systems Medicine, Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE), Bonn, Germany

Genomics and Immunoregulation, Life & Medical Sciences (LIMES) Institute, University of Bonn, Bonn, Germany

PRECISE Platform for Single Cell Genomics and Epigenomics at German Center for Neurodegenerative Diseases (DZNE) and the University of Bonn, Bonn, Germany

Pravin Shende Shobhaben Pratapbhai Patel School of Pharmacy and Technology Management, SVKM'S NMIMS, Vile Parle (West), Mumbai, India

Leonardo Seidi Shigueoka University of Campinas, São Paulo, Brazil

Zixin Shu Institute of Medical Intelligence, School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China

Sohaib Shuaat OMFS IMPATH Research Group, Department of Imaging and Pathology, University of Leuven and Oral & Maxillofacial Surgery, University Hospitals Leuven, KU Leuven, Leuven, Belgium

Meetali Sinha Computational Toxicology Facility, CSIR- Indian Institute of Toxicology Research, Lucknow, Uttar Pradesh, India

Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, Uttar Pradesh, India

Beril Sirmacek Smart Cities, School of Creative Technologies, Saxion University of Applied Sciences, Enschede, The Netherlands

Harpreet S. Sood Health Education England, London, UK

Viknesh Sounderajah Department of Surgery & Cancer, Imperial College London, London, UK

Institute of Global Health Innovation, Imperial College London, London, UK

Didem Stark Charité – Universitätsmedizin Berlin, Humboldt-Universität zu Berlin, Berlin, Germany

Department of Psychiatry and Psychotherapy, Berlin Institute of Health, Bernstein Center for Computational Neuroscience, Berlin, Germany

V. Stonawski UK Erlangen, Friedrich-Alexander University Erlangen-Nuremberg, Erlangen, Germany

Inga Strümke SimulaMet, Oslo, Norway

Robbert Struyven Moorfields Eye Hospital NHS Foundation Trust, London, UK

M. E. Suleiman University of Sydney, Sydney, NSW, Australia

Bowen Tang MindRank AI Ltd., Hangzhou, Zhejiang, China

Xavier Tannier Sorbonne Université, Inserm, Université Sorbonne Paris Nord, Laboratoire d’Informatique Médicale et d’Ingénierie des Connaissances pour la e-Santé, LIMICS, Paris, France

David Taylor The Patients Association, Harrow, UK

RSM Digital Health Council, Royal Society of Medicine, London, UK

Imperial College London, London, UK

James A. Taylor Google Health, London, UK

Bart M. ter Haar Romeny Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, the Netherlands

Vajira Thambawita SimulaMet, Oslo, Norway

Department of Computer Science, Oslo Metropolitan University, Oslo, Norway

Haoyu Tian Institute of Medical Intelligence, School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China

Daniel Shu Wei Ting Duke-NUS Medical School, National University of Singapore, Singapore, Singapore

Moorfields Eye Hospital, London, UK

Singapore Eye Research Institute, Singapore National Eye Center, Singapore, Singapore

Neil S. Tolley Imperial College London, London, UK

Imperial College NHS Healthcare Trust, London, UK

Marko Topalovic Laboratory of Respiratory Diseases and Thoracic Surgery, Department of Chronic Diseases, Metabolism and Ageing, Katholieke Universiteit Leuven, Leuven, Belgium

ArtiQ NV, Leuven, Belgium

Pablo Torres Instituto Universitario de Matemática Pura y Aplicada, Universitat Politècnica de Valencia, Valencia, Spain

Laurent Tournois BioSilicium, Riom, France

UMR 8045 BABEL, University of Paris, Paris, France

Michael B. Tradewell Department of Urology, University of Miami Miller School of Medicine, Miami, FL, USA

Andrew Trask OpenMined, Oxford, UK

Department of Computer Science, University of Oxford, Oxford, UK

Darren Treanor Leeds Teaching Hospitals NHS Trust, Leeds, UK

University of Leeds, Leeds, UK

Department of Clinical Pathology and Department of Clinical and Experimental Medicine, Linköping University, Linköping, Sweden

Centre for Medical Image Science and Visualization (CMIV), Linköping University, Linköping, Sweden

Victor Troussel IRIS Institut de Recherche Interdisciplinaire sur les enjeux Sociaux, UMR8156 CNRS – U997 Inserm – EHESS – Université Sorbonne Paris Nord, Paris, France

Department of Forensic and Social Medicine, AP-HP, Jean Verdier Hospital, Bondy, France

Samyakh Tukra Department of Surgery and Cancer, Imperial College London, London, UK

Eiichiro Uchino Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto University, Kyoto, Japan

Department of Nephrology, Graduate School of Medicine, Kyoto University, Kyoto, Japan

L. Uhlenberg Digital Health, Friedrich-Alexander University Erlangen-Nuremberg, Erlangen, Germany

Joost van der Putten Eindhoven University of Technology, VCA Group, Eindhoven, The Netherlands

Fons van der Sommen Eindhoven University of Technology, VCA Group, Eindhoven, The Netherlands

Josep Vehí Modelling, Identification and Control Engineering Laboratory (MICELab), Institut d'Informatica i Aplicacions, Universitat de Girona, Girona, Spain

Centro de Investigación Biomédica en Red de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM), Barcelona, Spain

Paola Velardi Department of Computer Science, Sapienza University of Rome, Rome, Italy

Alfredo Vellido Universitat Politècnica de Catalunya, Barcelona, Spain

Intelligent Data Science and Artificial Intelligence (IDEAI-UPC) Research Center, Barcelona, Spain

Ricardo Vinuesa FLOW, Engineering Mechanics, KTH Royal Institute of Technology, Stockholm, Sweden

Xaviour Walker Department of Medicine, University of Otago, Dunedin, New Zealand

Zhaoran Wang Duke-NUS Medical School, National University of Singapore, Singapore, Singapore

Stefanie Warnat-Herresthal Systems Medicine, Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE), Bonn, Germany

Genomics and Immunoregulation, Life & Medical Sciences (LIMES) Institute, University of Bonn, Bonn, Germany

Debra L. Waters Department of Medicine, University of Otago, Dunedin, New Zealand

Freida Werdiger Melbourne Brain Centre at Royal Melbourne Hospital, Melbourne, VIC, Australia

Department of Medicine, University of Melbourne, Melbourne, VIC, Australia

Paulina Werner Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden

Lukas Wisgrill Division of Neonatology, Pediatric Intensive Care and Neuropediatrics, Comprehensive Center for Pediatrics, Department of Pediatrics and Adolescent Medicine, Medical University of Vienna, Vienna, Austria

Nicolai Wohns Department of Philosophy, University of Washington, Seattle, WA, USA

Wojciech Wojnowski Department of Analytical Chemistry, Gdańsk University of Technology, Gdańsk, Poland

Tien Yin Wong Duke-NUS Medical School, National University of Singapore, Singapore, Singapore

Singapore Eye Research Institute, Singapore National Eye Center, Singapore, Singapore

Alex Wright Leeds Teaching Hospitals NHS Trust, Leeds, UK
University of Leeds, Leeds, UK

Artur Yakimovich Artificial Intelligence for Life Sciences CIC, London, UK

Dengying Yan Institute of Medical Intelligence, School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China

Guang Yang Cardiovascular Research Centre, Royal Brompton Hospital, London, UK

National Heart and Lung Institute, Imperial College London, London, UK

Yuxia Yang Institute of Medical Intelligence, School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China

Moi Hoon Yap Manchester Metropolitan University, Manchester, UK

Okuno Yasushi Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto University, Kyoto, Japan

Jong Chul Ye Department Bio and Brain Engineering & Department Mathematical Sciences, Korea Advanced Institute of Science & Technology (KAIST), Daejeon, Republic of Korea

Seng Chan You Department of Preventive Medicine, Yonsei University College of Medicine, Seoul, South Korea

Anastasia Young-Gough Preventive and Social Medicine, Dunedin School of Medicine, Dunedin, New Zealand

Ryan Carl Yu Harvard Business School, Cambridge, MA, USA

Hector Zenil Oxford Immune Algorithmics Ltd, Reading, UK

Alan Turing Institute, London, UK

Algorithmic Dynamics Lab, Unit of Computational Medicine, Karolinska Institute, Stockholm, Sweden

R. Zhang Digital Health, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany

Alex Zhavoronkov Insilico Medicine, Hong Kong Science and Technology Park, Hong Kong, China

Three Exchange Square, The Landmark, Deep Longevity, Inc, Hong Kong,
China

Buck Institute for Research on Aging, Novato, CA, USA

Keni Zheng Division of Physics, Engineering, Mathematics and Computer
Science, Delaware State University, Dover, DE, USA

Weihong Zhou Health Management Centre, Drum Tower Hospital Affiliated
to Nanjing University Medical School, Nanjing, China

Xuezhong Zhou Institute of Medical Intelligence, School of Computer and
Information Technology, Beijing Jiaotong University, Beijing, China

Alexander Ziller Institute for Diagnostic and Interventional Radiology,
School of Medicine, Technical, University of Munich, Munich, Germany

Institute for Artificial Intelligence in Medicine and Healthcare, School of
Medicine and Department of Informatics, Technical University of Munich,
Munich, Germany

OpenMined, Oxford, UK

Reyer Zwiggelaar Aberystwyth University, Aberystwyth, UK

Part I



Basic Concepts of Artificial Intelligence: Primed for Clinicians

1

Niklas Lidströmer, Federica Aresu, and Hutan Ashrafiyan

Contents

Introduction	4
AI, Machine Learning, and Deep Learning per Definition	4
A Brief History of AI	5
Rising Demand for AI	6
AI Applications	6
AI Staging	7
AI Programming Languages	7
Machine Learning	8
Types of Machine Learning	10
Machine Learning Problem Solutions	11
Limitations of Machine Learning	16
Introduction of Deep Learning	16
Single-Layer Perceptrons	17
Natural Language Processing	19
Conclusions	19
References	20

N. Lidströmer (✉)

Department of Women's and Children's Health, Karolinska

Institutet, Stockholm, Sweden

e-mail: niklas.lidstromer@ki.se

F. Aresu

KTH Royal Institute of Technology, Stockholm, Sweden

e-mail: aresu@kth.se

H. Ashrafiyan

Department of Surgery and Cancer, Imperial College

London NHS Trust, London, UK

Institute of Global Health Innovation, Imperial College

London, London, UK

Hamlyn Centre for Robotics and Artificial Intelligence,

Department of Surgery and Cancer, Imperial College

London, London, UK

e-mail: h.ashrafiyan@imperial.ac.uk

Abstract

With the urgent need for automatized algorithm applications to an ever-increasing amount of data and a further decrease of the chances of human errors on crucial tasks, artificial intelligence algorithms were introduced.

An expansive demand of AI applications in varying fields led to the development of specifically designed ad hoc algorithms with the role of better estimating (by learning) solutions to the problems.

The boost of AI in healthcare right now is a consequence of two things – the availability of big data and better processors, able to train and execute algorithmic tasks, i.e., implementations of these algorithms with neural networks.

It will soon be vital for medical students to grasp the principles of AI. The purpose of this major reference textbook on AI in medicine, of which this chapter is the base level introduction, is to become the greatest standard reference work. No area of medicine, preclinical or clinical, will escape the profound effects of AI: the whole healthcare domain will be reshaped thoroughly.

Keywords

Artificial Intelligence · Medicine · Basic Concepts · Introduction · Classification · Healthcare · Machine Learning · Deep Learning · Neural Networks · Programming

Introduction

Let's look at the AI landscape. Today we dwell on the plains of AI. It will heavily influence medicine, yet relatively few healthcare professionals still have a good understanding of the concept to come. Therefore, this chapter contains the scientific fundamental principles of AI – primed for clinicians. This serves as the path leading up to the basecamp location, where the evolution of deep medicine is elaborated. Hereafter, the path leads up to the mountainous high-level plateau of section III and its

presentation of AIM for the many medical specialties. These are the purposes of these first upward-sloping components of this book, which contents should be natural parts of the understanding of AIM by medical students and healthcare professionals of the future, in general.

After these general words on AI, a brief history of AI is presented, and an explanation to why it has become so famous right now. Then we must define exactly what AI is, its different stages, what language is best fitted for AIM, and how to properly communicate with programming experts. This part also contains a short intro to machine learning, with some medical demos. It will help to better grasp the contents of this large book, and understand the limitations of machines, and why machine learning is a necessity for the upcoming revolution of deep learning – a cornerstone in deep medicine. Other concepts in this first pedagogic book part are deep neural networks, natural language processing, and the practical implementation of the latter into medicine and in particular into AIM.

AI, Machine Learning, and Deep Learning per Definition

What exactly is AI? How was it defined when it first emerged as a term in 1956? According to John McCarty it is the *science and engineering of making intelligent machines*.

Hence, it is the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision making, and translation between languages.

AI is a technique of getting machines to work, act, or behave like humans. Recently we have started to realize this – robots and, in general, measurements-processing machines are used in many fields, such as healthcare, marketing, robotics, stock markets, business analytics, transportation, surveillance, etc.

Machine Learning (ML) is a subset of AI techniques and refers to a computer program that can *learn* how to produce behavior not explicitly programmed by the program's author.

This behavior is learned based on the data given, by minimizing the error between the current action and the ideal one, and a feedback mechanism (the famous “error backpropagation”) that uses it to lead the machine to self-recognize, and internally improve its performances.

Deep Learning (DL) is a subset of machine learning, which uses multilayer neural networks to solve complex problems. They exploit the processing of contextual data. Deep learning techniques are, therefore, more suitable for learning unstructured data.

A Brief History of AI

The history of AI is very old – it goes far back to antiquity and Greek mythology. Talos was a giant animated bronze warrior, programmed to guard the island of Crete, created by Hephaestus, to throw rocks at nearing enemy ships. But there are several myths about mechanical men and automata throughout human history. These were generally well thought of [1].

Modern medicine is intrinsically connected to advanced mathematical analysis, now requiring computers with fast processors. This has led to new diagnostics, pharmacovigilance, therapies, robot-assisted interventions, epidemiological data mining, and synthesis of evidence-based medicine and decision support.

Also, mathematics in medicine has a long history. Leiden professor of medicine *Archibald Pitcairne* (1652–1713) developed a theory of *iatro-mathematics* (medicine and mathematics) and can be regarded as the father of mathematical medicine [2]. During his career he stood under documented influence by mathematicians such as David Gregory (1659–1708) and Isaac Newton (1642–1727), whose *Philosophiae Naturalis Principia Mathematica*, heavily influenced the formation of a *Newtonian medicine*, and the formalized concept of iatro-mathematics, with lectures in Leiden during the early 1690s [3].

In 1920, *The Lancet* published an article of the topic of iatro-mathematics [4], where it was noted that the “rapprochement of medicine and mathematics is incomplete.” However, it was concluded

that the two sciences would “eventually become a firm alliance.”

In 1950, *Alan Turing* published a landmark paper, in which he speculates about the possibility of creating machines that think. He created a test known as the Turing test, aimed at determining whether a computer can think like a human being. He concluded that *thinking* is hard to define. *If a machine can carry out a conversation, which is indistinguishable from a conversation with a human being, then it would be reasonable to say the machine is thinking.* The machine would pass the Turing test. No machine until this day has achieved this result. The Turing test stands out as the first contribution to the philosophy of artificial intelligence.

Followed by this contribution came the era of “game AI” during the 1950s. In 1951 with the usage of the Ferranti Mark 1 machine, at the University of Manchester, the computer scientist Christopher Strachey wrote a checkers program, and Dietrich Prinz wrote one for chess at around the same time.

These were the first attempts to let computers play games such as chess, and compete with humans.

This was followed by probably the most important year in AI history. In 1956, the concept of AI was coined for the first time at the Dartmouth Conference by Professor John McCarthy. In 1959 the first AI laboratory was established, marking the next coming period as the AI research era. The lab is called MIT, and is still in operation and famous.

In 1960 the first AI robot was implemented into the GM assembly line, and the first chatbot ELIZA was invented. This is the great grandmother of Siri and Alexa. Soon hereafter came the very well-known IBM Deep Blue, which in 1997 beat the world champion Garry Kasparov in a game of chess. This is regarded as maybe one of the first great accomplishments of AI.

In 2005 at the DARPA Grand Challenge, the racing team of Stanford University participated with Stanley, an autonomous car, which won the race.

In 2011 IBM’s answering system Watson defeated two of the greatest Jeopardy! Champions Brad Rutter and Ken Jennings.

AI started as a hypothetical situation, evolved and is now the most important technology in today's world. AI has presented an exponential growth of its potential. Wherever we look around us AI deep learning or machine learning power many things.

Today AI dominates Knowledge Base, Expert Systems, Deep Learning, Computer Vision and Image Processing, Machine Learning and Natural Language Processing, etc.

Rising Demand for AI

A common query is how AI could have existed for over half a century and suddenly now appears as a “sudden” hype and attracting all attention? The main reasons for the present demand for AI are:

- *First of all*, we have more computational power now to train deep learning models, and one of the most important contributions to these technology improvements are graphic processors units(GPUs) for massively parallel processing at low cost. The improved computational power makes it possible to broadly and globally implement AI.
- *Secondly*, the enormous amount of data presently at hand. The data generated is at an immeasurable pace. The sources are vast networks of social media, IoT devices, mail, conversations, photos, medical imaging, and numerous other places. Hence, there is a demand for a method or solution to process this overload of data, in order to give us insight from it and let businesses grow as a result. This process is AI in essence. AI is trained on large datasets, big data, to help us make smart decisions, to classify objects in images, etc. These processes enable us to act more efficiently.
- Next up, we have far better algorithms. These are much more effective and based on the concept of neural networks, i.e., the deep learning architecture. All this enables quicker and more accurate computations.
- And last, but not least, governments, venture capitalists, tech giants, and start-ups are now all focused on AI and pour in investments. For instance, companies in the FAANG group (Facebook, Apple, Amazon, Netflix, Google),

Microsoft, and most car manufacturers, and a long list of major tech companies, are deeply investing in AI. The consensus among all mentioned instances is that AI is the way of the future.

For more details on the importance of the AI in medicine, please see the chapter “On the Importance of AIM,” by Dr Katarina Gospic et al.

AI Applications

Now that the definition is set, let us briefly mention a few significant applications, to highlight the importance of AI. The most famous is likely Google’s predictive search engine. It is in global use; whenever a person starts to type, Google makes immediate suggestions the user could use. This is AI in action literally. The predictability is based on collected data from individuals, browser usage, location, personal info, age, gender, and many more. Behind this guessing, there are many layers of natural language processing, deep learning, and machine learning.

Another striking application is in the financial field – J.P. Morgan uses the Chase’s Contract Intelligence Platform (COiN), which uses AI, machine learning and image recognition software to analyze legal documents. In this way the company avoided manually reviewing ca. 12,000 documents, which took more than 36,000 h. But when this monotonous task was replaced by the AI machine it took a few hours.

Since this book is focused on AI in medicine, let us mention some applications for healthcare. For instance, the Watson computer from IBM. Watson uses natural language processing, evidence-based learning capacities, and hypothesis generation, which has contributed to clinical decision support systems and contributed to AI in healthcare, and is today in use in an increasing number of medical specialties [5]. Medical doctors can pose questions to Watson, entering clinical facts such as symptoms, medications, and heredity, and Watson can then mine the patient data, examine available data, form a hypothesis, and finally provide a list of individualized confidence-scored suggestions [6].

Watson’s data sources encompass research articles, clinical studies, treatment guidelines, and electronic health record information [5]. In should

though be noted that not been directly involved into the medical diagnosing, it has only assisted with treatment alternatives for already readily diagnosed patients [7].

During the last 10 years Watson has partnered with a long range of organizations, companies, and universities, e.g., Columbia University, University of Maryland [8], Memorial Sloan-Kettering Cancer Center [9], MD Anderson Cancer Center, Manipal Hospital, Cleveland Clinic, and Case Western Reserve University [10].

The FAANG group and other large corporations have started AI initiatives within health; Facebook (*Preventative Health*, 2019–), Microsoft (*Health Vault*, 2011–2019, *Apple* (Health, 2014), *Amazon* (Amazon Care, 2018–), Google (*Google Health* 2006–2012, and with *DeepMind* 2018–) [11].

For instance, *DeepMind Health* collaborates with Moorfields Eye Hospital and are developing AI applications for healthcare, especially eye scanning [12], and with University College London Hospital aiming to develop an algorithm to differentiate between healthy and cancerous tissues in the head and neck region [13].

In global large networks, such as Facebook, AI is used in, e.g., face verification, both as password and auto-tagging of friends, and to personalize advertising systems using neural networks, machine learning, and deep learning concepts. Many people are not aware of how much AI they use on a daily basis in their lives – all social media platforms, e.g., Facebook, Instagram, Twitter, LinkedIn, heavily rely on AI. Through the 2016 US elections, political ads using social media are called to the spotlight. Specifically, the controversy of targeting users to obtain their personal information and determining what advertisement would persuade those electors.

Twitter uses AI to identify hate speech and terroristic language in tweets. In this way they detected ca. 300,000 terror-linked accounts. These nonhuman AI machines found 95% [14].

Also, virtual assistants such as Alexa and Siri have entered the market, and quite recently also Google Duplex, which responds to calls, can book appointments and with a human touch, making it sound realistic.

Some other examples of AI are Tesla's and many other car manufacturers' experiments with self-

driving cars and even taxis without a driver, based on a range of AI implementations. Also, Netflix uses AI and pattern recognition to make personal film recommendations. Gmail uses a similar principle to automatically sort incoming letters in mailboxes, spam filtering, etc. The latter uses buzzwords that are common in spam, e.g., full refund, lottery, etc., and then directs them to the spam compartment.

AI Staging

There are three main stages of AI

- **Narrow AI**, also known as weak AI. This can only be applied to specific tasks. Most applications today belong to this group, e.g., Alexa – although sophisticated, all functions are within a narrowly defined function range. Other examples are self-driving cars, chess computers, AlphaGo.
- **General AI**, or strong AI. We have not reached this stage. Strong AI refers to machines being able to possess the ability to perform any intellectual task that a human can. Machines now have strong processing powers, but hitherto no sign of reasoning capacity; hence we're stuck in the weak AI stage. Not even AlphaGo Zero, which learned without human intervention, could be defined as strong AI.
- **Super AI**, refers to a stage when computers would surpass human capacities. This stage encompassed big data statistics, symbolic mathematics, number of faces, generative adversarial networks (GANs).

AI Programming Languages

There are several languages used in AI applications, and one of the most popular and well-known is Python, partly because of its simple and functional syntaxes, and also for the great number of libraries designed for Python (to implement Machine Learning algorithms in a straightforward manner), such as Keras and Tensorflow.

The Python advantages are related to its simplicity and their maintainability as well as the possibility to connect and integrate with files written in other programming languages. Problems of memory usage and not having multithreading are substantial Python disadvantages [15].

Python was created in 1989 by *Guido Rossum*, and is an interpreted, object-oriented high-level programming language with dynamic semantics. Hence, it is a high-level language (no concerns with low-level details, e.g., memory allocation), which is free and open-source. Python is portable, i.e., it is supported by many platforms, e.g., Linux, iOS Mac, Windows PC, FreeBSD, Solaris, OS/2, Amiga, AS/400, BeOS, OS/390, PlayStation, Windows CE, etc.

Python supports various programming paradigms, i.e., both object-oriented, and procedure-oriented programming, and is extensible, i.e., it can invoke C and C++ libraries, and can integrate with a multitude of other languages, such as Java and .NET products. Python is the most rapid gainer in AI, with a huge momentum. Its use is ubiquitous to create AI algorithms, machine learning, IoT projects, etc. With Python, the developer doesn't need to code very much, because there are ready-made packages, with algorithms. For instance, *PiBrain* (for Machine Learning), *NumPy* (scientific computing, Pandas, etc.) can be implemented, and a vast range of libraries.

Apart from Python, another popular programming language used mainly for statistical tasks is called R. This language well-performs in analysis and manipulation of incoming data for statistical purposes. R is well known for its publication-quality plots and its compatibility with other programming languages. However, R is less suitable for handling big data analysis tasks for its consuming memory characteristic compared to Python and its speed in other programming languages.

R is almost as easy as Python to learn. Both languages are very similar to English in syntax and construction, hence they belong to the easiest to master. They both have an enormous number of libraries to provide all thinkable predefined algorithms, statistical models, data scientific inputs, AI, machine learning with algorithms, NLP, etc.

Moreover, Java is also used in AI, especially for artificial neural networks and genetic programming. Here Java has its benefits with, e.g., simple packaging and debugging, user interaction, and functionality for mega-project scalability and graphics. The latter is one of the outstanding assets of Java with its standard interface and graphics' toolkit – the graphical presentation is,

of course, a vital part of AI, and which will be seen in AIM, especially when applications are directed toward patients and students. Java provides better managing tools of garbage and provides multi-threading, differently from Python.

Another alternative language is *Lisp*, less known, but the most ancient and perhaps best-adapted language for AI development. Lisp goes all back to the origins of AI, and was introduced by John McCarty in the late 1950s, and can process symbolic information, can prototype, create dynamic objects, automatic garbbing, and is deemed easy by developers. Though, nearly all of its excellent features have migrated into many other languages. It is the Sanskrit, Swahili, or Latin of AI languages. The latter are more effective, have better packaging, etc.

SWI Prolog is a language, which is relevant in AIM, since it is often used in knowledge base and expert systems – it has features such as pattern matching, freebase data structuring, and automatic backtracking. This gives a strong and flexible framework for programming, which makes it frequently used in AIM.

Other languages worth mentioning are *C++*, *SaaS*, *JavaScript*, *MATLAB*, and *Julia*. All of these can be used for AI.

Machine Learning

Machine Learning (ML) is one of the most important instruments in AI. It is a way of feeding data into a machine, so it can make its own decisions. The need for ML is as old as the technical revolution, which has generated immeasurable loads of data. In research we generate over 2.5 quintillion bytes of data per day. In 2020 an estimated 1.7 MB of data was created every second for every person on earth. With this vast amount, models can be created to study and analyze complex data, insights and more precise results can hence be delivered. In the last 2 years alone, the astonishing 90% of all the world's data has been created. At the end of 2020, 44 zettabytes (10^{21}) made up the entire digital universe. 2.5 quintillion (10^{18}) bytes are created by humanity every day. It is estimated 463 exabytes (10^{18}) of data will be generated every day by

humans in 2025 [16]. The rememberable fact is also that we have decided to store this forever, costing us substantial energy.

In a worst-case scenario, computer technology could use as much as 51% of global electricity in 2030. This will happen if not enough improvement in electricity efficiency of wireless access networks and fixed access networks/data centers is possible. However, until 2030, globally generated renewable electricity is likely to exceed the electricity demand of all networks and data centers. Nevertheless, a recent investigation suggests, for the worst-case scenario, that electricity for computer technology usage could contribute up to 23% of the globally released greenhouse gas emissions in 2030 [17].

With machine learning in the finance sector, a wide range of profitable opportunities and avoidable risks can be identified. An understatement is of course that the equivalence will enter the medical and healthcare domains. The simple foundation of all machine learning, and of AI, is *data* per se. Data is the solution, we just need to know how to handle it, and ways to do this are ML, deep learning, and AI.

In medicine the needs for machine learning are associated with

- Increased *data generation* from electronic health records (EHRs), digitalization
- Improvements in *decision making*, risk prediction
- Uncovering of *patterns* and trends in data, finding hidden patterns, key data extraction
- Building statistical models, time saving
- Solving of complex problems for humans



An illustration of *Machine Learning* (ML) [18]

When *Arthur Samuel* coined Machine Learning in 1959, the definition was “*a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its tasks in T, as measured by P, improves with experience E.*”

In other words:

Data > Training the Machine
 > Building a Model > Predicting Outcome

Machine learning (ML) in all essence, is a subset of AI, which provides machines the ability to learn automatically and improve from data pattern analysis. The ML has been programmed in a way that it can adjust its parameters. Some more definitions of terms used in this textbook are the following:

Algorithms – Set of rules and statistical techniques to learn patterns in data, mapping all decisions that a model can take.

Model – A mathematical equation, i.e., a computation or a formula, which is the result of an algorithm that takes some values as input and produces some values as output.

Predictor feature – Variable of the data that helps predict the output, e.g., a physical factor in patients, where a gene, height, lab value, vital sign, or weight can predict a symptom or diagnosis.

Response Variable – The feature, output variable, or target variable that will be predicted.

Training data – The data, which is used to train the ML model.

Testing data – Unseen data used to test the ML model after the training procedure.

The *machine learning process* involves the construction of a Predictive Model, which is then used to identify a solution to a Problem Statement. Hence the process of ML involves several steps:

Definition of the objective > Data gathering
 > Data preparation > Data exploration
 > Model construction built on training
 > Model testing > Predictions

AIM ML case example:

- **Definition of Objective**

Formulation of the problem idea: e.g., to predict if an infected patient will recover, yes or no.

Q: What target feature: e.g., temperature, blood pressure, or other specific symptoms

Q: What input data: medical records, vital signs history, lab results, etc.

Q: What kind of problem: Binary classification? Clustering? Regression problem?

Hence, at step 1 we define how we will solve the problem, what kind of data we need, and what we are trying to predict, what information is needed to predict.

- **Data Gathering**

Data such as temperature curves, lab values, vital signs, physical examinations, massive loads of previous comparable patients EHRs etc., and where can we get this data – gathered manually or scrapping from the web, from EHRs, from other sources? This is the most time-consuming element of the ML process. We create the dataset.

For ML exercises there are a lot of training datasets online, e.g., Kaggle (<https://www.kaggle.com/>) and Grand Challenge (<https://grand-challenge.org/challenges/>), where a whole range of sets can be downloaded, having numerous different themes; weather forecasts, economic forecasts, etc. Hence, developers can skip the time-consuming data gathering step, and just download the dataset.

- **Curating**

Data preparation, or cleaning, involves erasing inconsistencies, e.g., missing factors or redundant information, erasing unnecessary data, format correction, and getting the data ready for analysis. Step 3 is likely the hardest step to perform. It is easy to bias the dataset if, e.g., one factor's values are missing more frequently. Any mistake will affect the result.

- **Exploratory Data Analysis**

This is the ML brainstorming step, which involves the understanding of patterns and trends in the data. Insights are concluded, such as correlations between variables. For instance, low blood pressure and alternating fever or other pre-septic signs are factors that

the output result will depend on. Exactly how, we have to find out in the patterns of the vast material, i.e., the correlations of such variables.

- **Building a Machine Learning Model**

A predictive model is created with ML algorithms, e.g., Linear Regression, Decision Trees, etc. This stage always starts with data splicing, into training data and testing data. The former is always used to build the model. Training data is usually 80% of the total.

In this AIM example, we are predicting the outcome of classification variables, also known as *categorical variables*, i.e., recovered patient yes or no. Two alternatives. In this case we can use linear regression, support vector machines, K nearest neighbor, or Naïve Bayes, etc. Which algorithmic model can be used depends on the problem statement, i.e., it depends on the task to solve. The methods suggested to better approach a specific type of task are the results of a continuous process of trial and error.

- **Model Evaluation and Optimization**

This step evaluates and tests the model, so it can be improved, parameters tuned, etc. A part of the testing dataset is used as a validation dataset. The accuracy and errors as a form of ML performance metrics are calculated.

- **Predictions**

The final outcome is used to make predictions about the given medical condition, and in this case the outcome will be a categorical one. First you get a probability, and based on a preset range, the clinician can decide whether the answer is yes or no.

Types of Machine Learning

There are basically three types of ML – supervised, unsupervised, and reinforcement learning.

In *supervised* learning, we train the machine with labeled data. In other words, the labels act like guides. In AIM, we may feed the machine with expert interpretations of, e.g., X-ray images, such as fracture or no fracture of a specific bone. We explicitly train the machine with these labels. This type is suitable for regression and

classification problems. The approach involves mapping labeled input to known output. Applicable algorithms include linear regression, logistic regression, Support Vector Machines, KNN, etc.

In *unsupervised* learning the machine trains with unlabeled data, which lets the machine work on unguided information. No labels are fed into the procedure. Here the machine classifies by itself; it will identify dissimilarities between an X-ray with a fracture and one without, by noticing another pixel pattern, other shadows, fracture lines, and other classifying features. It clusters into two groups in this example. This type is suitable for association and clustering problems. The approach is to understand patterns and discover output. Applicable unsupervised algorithms include Clustering using hierarchical clustering algorithms, k-means, and mixture models. Furthermore, Anomaly Detection algorithms such as k-nearest neighbor in medical image analysis are widely used algorithms together with Deep Neural Networks approaches such as Deep Belief Nets, autoencoders, and Self-Organizing Map (SOM).

In *reinforcement* learning, the agent is placed in an environment and learns how to behave inside these surroundings by performing certain actions and noting what kind of rewards are received for these. An example is how Robinson Crusoe started to adapt to the desolate island. He explored the environment and identified possible rewards and dangers. This type of learning is suitable for reward-based problem solutions. The approach is a trial and error method. Applicable algorithms include Q-learning, SARSA, etc.

Machine Learning Problem Solutions

ML problems can be classified into three types – regression, classification, and clustering problems.

Regression problems – Solved with supervised learning. The output will always be a continuous quantity, i.e., a variable with an infinite range of values, for instance, medical lab results. The body temperature can be 36.8, 36.9, 35.9, etc. The data is variable. The aim is to forecast or predict. It is solved with algorithms like linear regression. Regression is fitting a curve through data.

Classification problems – Solved with supervised learning. The output is always a categorical quantity. The main goal is to calculate the data category, for instance, classification into skeletal injury or undamaged structure. Or gender, two categorical outcomes. On ImageNet there are even 1000 classes. Solved with algorithms such as logistic regression, support vector machines, K nearest neighbor, etc.

Clustering problems – Solved with unsupervised learning. Assigns data points into clusters. The principal aim is to group similar entities into >2 clusters, based on feature similarity, for instance, to find hidden signs of severe disease. Hence, when the data is insufficient, and we don't know the output of, e.g., two categories, then we form clusters. Solved, e.g., by the algorithm of K-means.

Supervised Learning Algorithms

Below seven useful algorithms to a class of supervised learning are presented.

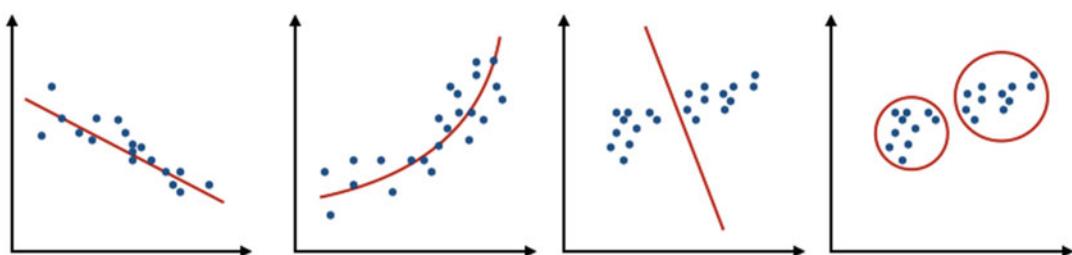


Fig. 1 Illustration of three machine learning problems; left to right: linear and nonlinear regression, classification, and clustering [19]

- *Linear Regression* – A method to predict the dependent variable that belongs to the y-axis, based on the values of independent variables along to the x-axis. The variables are continuous or discrete. Used for predictions of a continuous quantity. Hence the curve fitting the data is linear. The equation is:

$$Y = \beta_0 + \beta_1 X + e$$

Y Dependent variable

β_0 Y-intercept

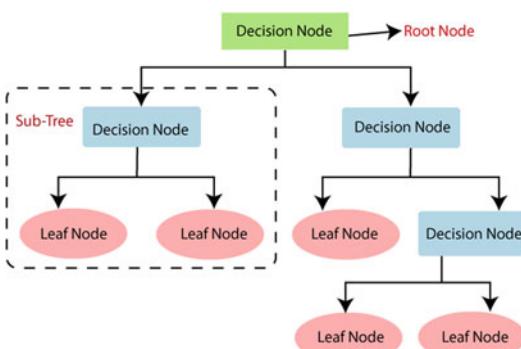
β_1 Slope

X Independent variable

e Error

For instance, variables with lab results can be imported in a CSV format into Python and then prepared, analyzed, and adequate algorithms applied.

- *Logistic Regression* – A method used to predict a dependent variable, from a dataset, when the dependent variable is categorical, and the outcome is 1 or 0, e.g., in orthopedic medical imaging, this would correspond to the presence of a bone fracture or no fracture. The function is sigmoid with values that go from 0 to 1. Afterward classification algorithms can be applied. See: https://en.wikipedia.org/wiki/Logistic_regression.
- *Decision Tree* – This is another classification algorithm, and looks like an inverted tree. It is a ML algorithm where each node signifies a *predictor variable* (aka feature), and the links between the nodes are decisions and at each branch there is a leaf, which stands for an outcome, or *response variable*.



An image showing an example of a decision tree [20]

Let us assume we have a large data set of ICU patients, and we would like to predict whether an infected patient faces the risk of a serious complication, e.g., sepsis, or not. Every step in the inverted tree represents a categorical classification step, i.e., a choice between two values in a binary tree. For instance, we use these short queries – “feverish or not,” then “normal blood pressure or not,” “normal pulse or not,” “shills or not,” “affected general condition or not,” etc. Each node that represents observations is directed through the branches, purely conjunctions of features, to the leaves also known as targets/labels. A common approach in triage or telephone consultation, where the nonmedical operator tries to decide whether a patient should seek the emergency room, see his/her GP the day or just wait.

The most significant clinical factor should be the initial root node, followed by internal nodes and then the terminal nodes or leaves lead to a suggested outcome. Branches are the answers, yes or no, 1 or 0, etc.

ID3 algorithm, *Iterative Dichotomizer 3* algorithm, is one of the most effective ways of building the tree within healthcare:

- Step 1: Selection of the best attribute (A) – e.g., affected general condition?
- Step 2: Assign A a decision variable for the root node – affected or unaffected patient?
- Step 3: For every value A can take, create a node descendant – if yes, if no, etc.
- Step 4: Add classification labels to the lead node
- Step 5: If the date results in correct classification, then stop.
- Step 6: If not, then iterate.

What best separates the data in a clinical decision tree is of course the most important factor first, and which classifies the patient group the best. During the construction, it is common to repeatedly try with different variables and analyze which is the most suitable. The two most important factors to consider here are *information gain* and *entropy*. The variable that best separates the data into the desired output classes is the variable with the highest information gain.

The entropy is a measure of *uncertainty* or impurity, which the data contains. The information gain (IG) signals how much *information* a specific variable or feature brings us, in regard to the final outcome, which can be concluded with this logic:

E.g., has the patient had a fever?

$p(\text{yes}) = \text{No. of yes outcomes in the parent node} / \text{total number of outcomes}$

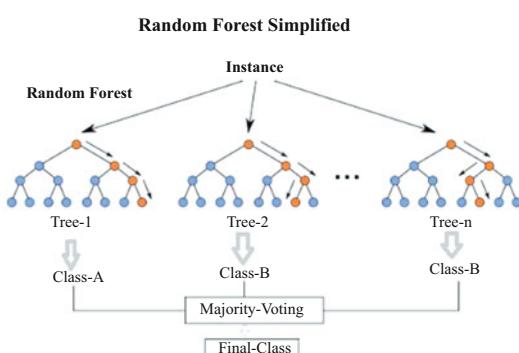
The entropy can be calculated in a similar manner:

$$\begin{aligned}\text{Entropy}_{\text{parent}} &= -p_{\text{yes}} \log_2(p_{\text{yes}}) \\ &\quad + p_{\text{no}} \log_2(p_{\text{no}})\end{aligned}$$

In this way, the variables' data splitting capacities are calculated to deem their suitability in a given place of the decision tree:

$$\begin{aligned}\text{Information Gain} &= \text{Entropy}_{\text{parent}} \\ &\quad - [\text{weighted average}] \\ &\quad * \text{Entropy}_{\text{children}}\end{aligned}$$

- *Random Forest* – Builds multiple decision trees (a forest) and fuses them and achieves a more precise and stable prediction. The forest gives more accuracy, avoids overfitting (when the models learn also the noise or disturbance and takes this into the model, which negatively affects the models' ability to predict from new data), and provides bagging, which means multiple trees test the data. This is suitable for more complex medical situations, e.g., a patient with multiple symptoms. From huge data sets of medical records, then bootstrapping can be executed in a row of circuits to make predictions.



An illustration of a *Random Forest* [21]

With the bootstrapped dataset we build a decision tree, starting at the root node, for which the best attribute is used to split the dataset. For each of the upcoming branch nodes, the process is repeated. For each step, the best attributes are chosen. The iteration is operated hundreds of times, and hence a forest of trees arises. The bootstrap dataset is used multiple times.

Finally, the prediction stage is reached. If we intend to predict a medical condition in a patient, we run the data through the forest of decision trees, and after all trees have been used it is presented which classification the majority of trees have voted for.

The model can be evaluated with a part of the dataset, which was not bootstrapped, in the so-called out-of-bag dataset.

- *Naïve Bayes Classifier* – A supervised classification algorithm, which is based on Bayes' Theorem, which solves classification problems with a probabilistic approach. The main idea is that the predictor variables in an ML model are intrinsically considered independent of each other. The concept is called naïve, since it does not consider any correlations between the variables. In medical real-world problems, there are often exactly such correlations anyway, but those are disregarded in this model.

The mathematics building up this model calculates the probability for an event to occur based on events in the past:

$P(A|B)$ – the conditional probability of the event A happening, given event B

$P(A)$ – the likelihood of event A happening

$P(B)$ – and of event B happening

$P(B|A)$ – the conditional probability of the event B happening, given event A

which gives

$$P(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

In a medical example this model could be used to classify, e.g., the likely infectious agent, based

on its diagnostic features – lab tests, microscopy, X-ray, quick tests, clinical signs, etc.

Infectious agent	One lab test	Microscopy	Quick test	Clinical signs
Infection type 1	4500/5000	0	0	5000/ 5000
Infection type 2	500/5000	5000/5000	4000/ 5000	0
Infection type 3	5000/5000	0	1000/ 5000	500/ 5000

Above, we have 15,000 patients, which can be divided into three groups of 5000 with one specific infection per group. Let's say we are given the following observation then:

	One lab test	Microscopy	Quick test	Clinical signs
Observation	Yes	No	Yes	No

To predict whether the infection is of a certain type, Naïve Bayes can be used. P is the probability. H is the hypothesis. C is a specific class. C1, class 1, etc.

$$P(H|\text{Multiple Evidences}) = P(C1) | H) * P(C2|H) \dots \\ * p(Cn|H) * p(H)/P(\text{Multiple Evidences})$$

All of the above alternatives are calculated, by inserting them separately, hence we receive the probability of which one is the likely infection in the observation. The conditional probability is delivered. There is a plethora of educational and clinically relevant examples online [22].

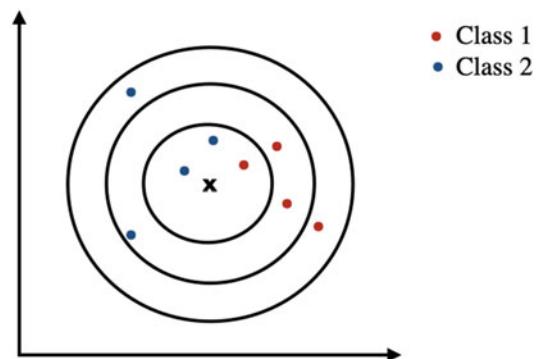
- *K-Nearest Neighbor* – KNN is a supervised learning algorithm, also used in unsupervised learning problems, which classifies a new data point into a target class, led by its distance to its closest neighbor points. For instance, this model is used to present medical images to the model, which classifies it into, e.g., if a skin mole is a melanoma, or a benign nevus.

KNN is a simple and easily applied ML algorithm of both classification (mainly) and regression type. It is nonparametric, i.e., it has no assumption, which is the case with Naïve Bayes, which assumes there are no relations between the variables. It is also a *lazy algorithm*, meaning it

remembers the training set, instead of learning a discriminative function. The most important feature is that it is based on feature similarity with neighboring data points.

The K value stands for the number of nearest neighbors, within a radius defined by, e.g., Euclidean or Manhattan distances – there are many possible distance metrics. When determining the nature of a novel data point, the number of members of preexisting class categories as neighbors are crucial. If a new data point would have one data point within its radius, which is of type A and two out of B, then it would be assigned the type B, if K were set to 3 (to include the three closest neighbors). If more neighbors are enclosed, then the type may change. The adequate K value can be calculated with, e.g., the elbow method, see below.

The data points (y, x) in a diagram are separated with a length, which is calculated with the Euclidean distance. Its simple equation is used by KNN to check the closeness of a new data point from its closest neighbors.



Example of K-Nearest Neighbors (KNN). Calculating the Euclidean distance of each point to the unclassified gray point and defining the K number of neighbor points to consider, the classification occurs. For instance, if K = 3, then x will belong to class 2. The x is in the center of the circles, which indicate the distance to a point (the cross). [19]

- *Support Vector Machine* – This is a classification and regression algorithm, which separates data, with the use of *hyperplanes*. SVM studies labeled training data. Support Vector Regressors are useful in regression problems; otherwise SVMs are mainly used in classification problems. Nonlinear data can be classified by

SVM with the use of kernel tricks. Nonlinear means the data cannot be separated with a sole linear line, elaborated below.

SVM works by drawing a boundary, a hyperplane, between different classes, and hence separating them in the best way. Support vectors are the data points closest to the hyperplane. The boundary will be drawn based on info on the support vectors. The optimal hyperplane has a maximal distance, i.e., margin, to the support vectors. It is easy as long as the hyperplane separation is linear. If this is not the case nonlinear SVM is used. Here the Kernel trick transforms the data into another dimension, e.g., separating them on the z-axis instead, if a 2D-approach didn't allow separation with a straight line. In the 3D-space a clear hyperplane might then be visualized, if the data allow.

Python can easily demonstrate and run all of the mentioned algorithms mentioned in this introductory part of the book.

Unsupervised Learning Algorithms

The main aim of the unsupervised learning algorithm *K-means clustering* is to group similar data points into a cluster. The process classifies objects into a predefined number of groups, so they are as dissimilar as possible between the groups, and as similar as possible within the group.

Every cluster has a *centroid*, from which distances to the objects are calculated. Grouping is then based on minimum distance. When faced with new medical population material, we need to first guess how many clusters there may be, and then provide a centroid for all these. The algorithm then calculates the Euclidean distance of the points from each centroid and assigns the point to the most proximate cluster. Over and over the clusters can be recalculated and new points added, and these steps are repeated until the centroids become the average of the cluster, i.e., iteration until the centroid value doesn't change.

With the *elbow method* the most optimal K value for a given problem is calculated. First the *sum of squared errors* (SSE) is computed for some values of K. The definition of SSE is defined

as the sum of the squared distance between each data point, or member, of the cluster and its centroid. The method can be plotted as graph of the WCSS(x) value (within-cluster sums of squares), using this formula:

$$\text{WCSS}(k) = \sum_{j=1}^k \sum_{x_i \in \text{cluster } j} \|x_i - \bar{x}_j\|^2$$

Where \bar{x}_j is the sample mean in cluster j

Elbow method equation [23].

The *distortion* will decrease with increased number of clusters, i.e., with a higher K value. At one point the distortion tilts abruptly, like an elbow in the graph. In large AIM studies, it is critical to pick the best amount of clusters.

Reinforcement Learning

Reinforcement learning algorithms consist of two main components – agent and environment.

For instance, the RL agent learns from the environment by rewards or failures, like if exposed to a new computer game. The RL iterates the game until it masters it. Some main concepts:

Reward (R) – The instant return from the environment to feedback the last agent action

Policy (π) – The approach the agent uses to decide the next action

Value (V) – The expected long-term award with *discount*, in opposition to short-term reward

Action-value (Q) – Like Value, but includes an extra parameter, the *current action (A)*

The Reward Maximization Theory states that an RL agent must be trained in such a manner that it takes the best action, so that the reward is maximized.

Discount means escaping negative events. Discounting is measured with a *gamma* value between 0 and 1, with larger discount the smaller gamma value.

Two other very important terms are *exploration* and *exploitation trade-off*. The former is about exploring the environment, the latter about exploiting the environment.

Another concept is the *Markov's Decision Process*, which is the mathematical approach for mapping a solution in reinforcement learning, and includes the following parameters:

Set of actions, A
 Set of states, S
 Reward, R
 Policy, π
 Value, V

For example, this can be shown in the shortest path problem, i.e., to calculate the shortest path between two nodes, with the minimum possible cost.

Q-Learning algorithms are among the most important examples of reinforced learning, which can be used to reinforce, e.g., the pathways taken by an agent, which is taught to get out of a labyrinth with several rooms, if this is the set goal.

The memory learnt by the agent with experience is represented as a *Q matrix*, where the rows represent the current agent state and the columns the possible actions leading to the next state, which results in the final formula:

$$\begin{aligned} Q(\text{state}, \text{action}) &= R(\text{state}, \text{action}) \\ &+ \text{Gamma} * \text{Max}[Q(\text{next state}, \text{all actions})] \end{aligned}$$

The γ (gamma) parameter ranges as said 0 to 1, and if γ is closer to 0, the agent will consider immediate rewards, and if closer to 1, the agent will weigh future rewards greater.

Hence, γ closer to 0 leads to exploitation, while γ closer to 1 leads to exploration.

With these words I round off this simple mathematical review, and tilt our focus to other general considerations on AI, ML and DL, e.g., the definitions and their limitations, which come next.

Limitations of Machine Learning

ML is first of all not capable of handling high-dimensional data, where the input and output is very large, since this complex data consumes resources, and causes curse of dimensionality, as several new dimensions are added. Just going

from 1D, 2D, and to 3D demands quite new requirements. And in a real-life situation such as in AIM there can be thousands of dimensions.

One of the big challenges with traditional ML models is a process named feature recognition, involved in object recognition, handwriting recognitions, and natural language processing, which consume a lot of resources. In these types of tasks, the dataset size has a big impact on machine learning performances. It is necessary to have a larger dataset for learning such variety and complex data. Therefore, the usage of deep learning approaches, such as CNN with hundreds of millions of trainable weights, is required.

Introduction of Deep Learning

Deep learning (DL) models are capable of focusing by themselves on feature extraction with very little guidance from the programmer, and can solve especially the dimensionality problem – to avoid the curse of dimensionality [24]. The main idea is to imitate the structure of the brain. DLs learn by themselves automatically, and they can quickly identify the decisive factors.

For instance, DL can be used for facial recognition. If ML is used, all specific elements of a face must be separately defined; but with the neural networks of DL, they will automatically detect the key features of each individual.

Deep learning is a form of ML, which uses a computing model, hugely inspired by the cerebral structure, where dendrites receive signals to the neuron from other neurons, the neuronal cell bodies summarize all the inputs and the axons transmit these to other cells, by firing off at a certain threshold, etc.

In DL the dendrites are represented by artificial neurons, also called perceptrons. They hence receive input of data from multiple sources, and deliver an output. The input and output are exactly like predictor variables. Data fed to a perceptron will undergo different functions and transformations, and then give an output. The perceptrons are connected into an artificial neural network. DL is a collection of statistical ML techniques, which

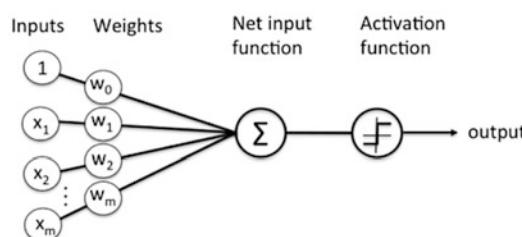
learn feature hierarchical structures, based on the concept of artificial neural networks.

In this neural network structure, there are different types of layers – the input layer receiving all the inputs, and the delivering output layer. Between these layers are numerous amounts of hidden layers. The number of perceptrons in each layer and the number of layers depend on the problem in question.

For image recognition in radiology, at first the high dimensional data is presented to the input layer, which in this case due to the massive load will contain multiple sublayers of artificial input neurons to digest the entire input. The output received from the first input layer contains extractable patterns, identifying the edges of the images, based on contrast levels. This data will be relayed to the first hidden layer, which will be able to identify certain clinically relevant features (or in facial recognitions, facial features), further transfer to the second hidden layer will be able to create a fuller picture (or the entire face). If the deeper recognition's entirety is sufficient (and no more layers needed), then the data is transferred to the output layer to deliver the classification.

Single-Layer Perceptrons

An Artificial Neural Network (ANN) with a single layer used for binary classification tasks is a linear model. The ANN, like the neuron, has a set of inputs, and each of these is dedicated to a certain weight, and it computes some specific functions on these weighted deliverances and produces an output. It separates the data into two separate classes (low or high).



Schematic of Rosenblatt's perceptron.

Principle of single-layer perceptron [25]

The perceptron above receives different inputs X_1, X_2, \dots, X_n , etc., and biases, and those are weighted accordingly to $W_1, W_2, \dots, W_{n1}, Y_2, \dots, Y_n$.

The word activation function alludes to its equivalent in the human cerebral cellular function with the action potential, where the neurons are activated at a certain threshold. It is mirrored here mathematically, when the function becomes saturated above a given threshold.

There are several types of activation functions, e.g., signum, sigmoid, tan, hedge, etc. All functions match the input to the respective output. The bias can shift the activation function in order to get an exact output.

As a medical perception analogy – should a person at the ER be sent to the ICU or not? There are three cardinal symptoms, X_{1-3} , and those have an importance or weight of W_{1-3} . If the symptom is present X attains the value of 1 and if not 0. W_1 is set to 6, W_2 to 2, and W_3 to 2. Then the firing threshold is considered. If the threshold is 5, it means the patient will be admitted if symptom 1 is present, even if the two others are not. If the threshold is 3, then either symptom 1 or the two other symptoms together will be sufficient for patient transfer.

X_1 – symptom 1 with W_1 of 6

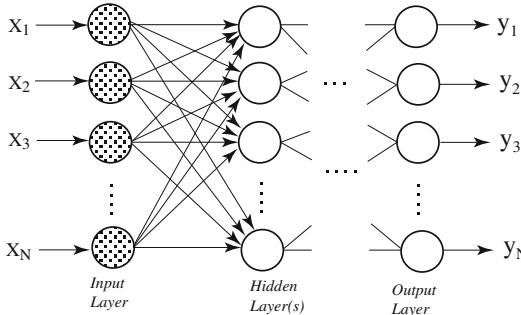
X_2 – symptom 2 with W_2 of 2

X_3 – symptom 3 with W_3 of 2

The limitations of perceptrons are associated with the absence of hidden layers, with the consequence it can only be applied to linear problems. When dealing with a nonlinearly separable data a single-layer approach is not feasible. A multilayer perceptron with back-propagation can though be used to solve non-linear problems.

Multilayer Perceptron – Artificial Neural Network

A multilayer perceptron has the same structure as a single layer perceptron, but with one or several hidden layers, and hence is considered a deep neural network.



A multilayer Artificial Neural Network (ANN) [26]

The weights between the layers of the deep network are the principal way in which long-term info is stored in artificial neural networks. The ANN learns new info by altering the weights or keeping them up to date. A set of new inputs passes through the first hidden layer, and then to the next as shown in the figure. Initially in an ANN the weights are assigned as random. With backpropagation, which is the quintessence of supervised neural network training, the weights of a neural network are fine-tuned. The tuning is based on the error rate obtained in the previous epoch (or iteration). Done properly this reduces the error rates and makes the model reliable by increasing its generalization. In order to have a correct output and a reduced error rate, the weightage needs to be adequate. The back-propagation is a way to train the perceptron by updating the weights.

The most common DL algorithm for supervised training of a multilayer perceptron is back-propagation. After the forward pass through the whole network, it is propagated backward, the

results updated with new weightage to minimize the error (compared to the wanted output). The procedure is iterated using gradient descent to do it faster and the errors/losses are reduced and the output optimized.

The limitation of a feed forward network can be exemplified with, e.g., image recognition. The network is exposed to a set of images, where the first photo presented will not really change the way it classifies the next photo. So the output at the time T is independent of the output at time T-1. This concept cannot be a logic approach if the T-1 information is necessary to understand the next information coming at T. Here we need a concept, which is the same as you apply, when you read this book – you need to understand parts I–III before you proceed to the systematic review in part IV. In an ANN the info at T + 1 has nothing to do with the info at T or T-1. Hence, an ANN cannot be used in situations where the output is based on previous outputs, e.g., for predictions of words in a sentence.

The solution to this problem is a *Recurrent Neural Network* (RNN), which is a type of ANN designed to recognize patterns in series of data, e.g., genomes, text, speech, handwriting, or numerical times series from medical sensors, or from stock markets, government agencies, etc. The applications are vast for RNNs. The prediction is based on past output.

Another concept is the *Convolutional Neural Network* (CNN), which is useful for image processing. A computer sees an image by breaking it down to three color channels, red-green-blue

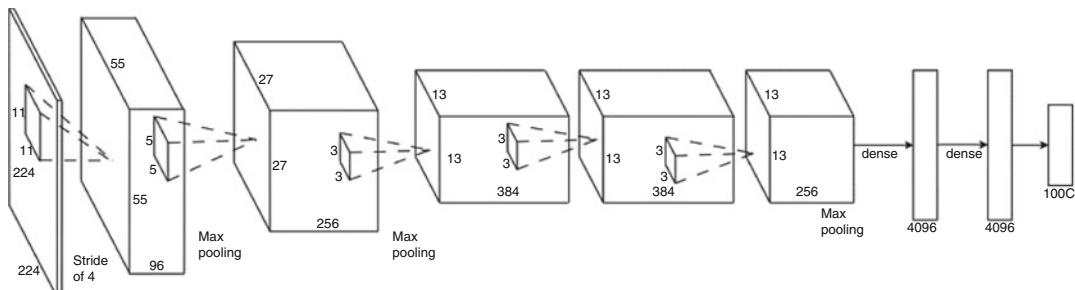


Fig. 2 A Convolutional Neural Network (CNN) architecture, with (fully connected) dense layer of 4096 units, derived from the last max-pool layer to the right (dimensionality change) [27]

(RGB), and reads the image with RGB values. Each of the channels is mapped with the pixels of the image. (The computer calculates the value of each pixel, and delivers the image size). In a CNN, a neuron will only be connected to a small portion of the previous layer. The CNNs are not fully connected.

Natural Language Processing

The main reason for the need of text mining and Natural Language Processing (NLP) is the stratospheric amount of data generated *every day*, and it is expected to grow. For instance:

Instagram	8.95 million photos posted/day
Twitter	500 million tweets/day
Facebook	3.2 billion likes and 350 million photos/day
Mail	300 billion emails sent/day

The vast majority of all data is completely unstructured. With the mining, structuring, and analyzing the data, huge scientific and economic values can be harvested. This is the essence of text mining – the process of deriving meaningful data out of natural language. NLP is a method used in text mining, and is the part of computer science and AI, which deals with human languages, and helps computers read human texts. A few applications of NLP and text mining are: auto-correction, auto-completion in searches, spam email detection, predictive typing, spell-checkers, and email classification. Sentimental analysis provides insights into the public or customer opinions in certain topics or products. Social media platforms use this analysis continuously. Chatbots are widely used in, e.g., customer services. Speech recognition in use by Siri, Cortana, and Alexa are all NLP applications. Machine translation is also an example of NLP, for instance, in use by Google Translate. Other NLP applications include information extraction, keyword search, and advertisement matching.

The *Generative Pre-trained Transformer* (GPT-3) is an autoregressive language model, which uses deep learning to produce human-like text. In the GPT series, it is the third-generation language prediction model, and was created by OpenAI in San Francisco [28]. The capacity of

GPT-3 is 175 billion machine learning parameters, and is the currently largest example of natural language procession (NLP) model [29].

A major concept in NLP is *tokenization*, which is the splitting process of the whole data (corpus) into smaller items (chunks). First it breaks a complex sentence into words, then analyses each single word's importance in the sentence and then translates. *Stemming* is the procedure of normalizing the word into its root or base or nominative state. The stemming algorithm provides this, with limitations though. In order to improve the accuracy *lemmatization* uses a large dictionary to group different inflected forms of a word, i.e., *lemma*.

Similar to stemming it labels a group of words into a common bundle. The output is then a usable proper word. Stemming may lead to an indiscriminate cutting of the word, so that the grammar or the understanding is lost, hence the introduction of lemmatization, which always delivers a word found in a dictionary. Lemmatization occurs after the morphological analysis.

Stop words are critical to applications – they can be removed with no loss of meaning, in order to focus on the *keywords*. The stop words will only decrease search accuracy.

Another concept, useful in natural language processing (NLP) is the *document-term matrix*, which is a mathematical matrix, which describes the frequency of terms occurring in a set of documents. Columns correspond to terms and rows to documents. Features can also be used besides terms [30].

Natural language processing has enormous potential for text mining of electronic health records and other medical applications, as will be elaborated in many other chapters of this reference textbook.

Conclusions

In conclusion, machine learning and deep learning algorithms have been widely implemented in various sectors with noticeable exponential application demand.

There will be no single area in preclinical medicine or clinical specialty, which will not be

profoundly affected by applications of artificial intelligence in medicine. AI will execute a thorough recast of all fields of healthcare.

The era of *mathematical medicine*, which can be said to have started with Archibald Pitcairne already in the seventeenth century [2], now with the emerging AI in medicine era, comes into full fruition.

This book encompasses the future potential of medicine and the benefits that can be unleashed when AI platforms can unlock the strengths of Big Data for Healthcare.

References

1. Sparkes B. The Red and The Black: studies in Greek pottery. Routledge; 1996. p. 124. ISBN 0-415-12661-4, ISBN 978-0-415-12661-8; two late fifth-century vase paintings depicting the death of Talos are discussed by Robertson M. The death of Talos. *J Hellenic Stud* 1977;97:159f.
2. Ashrafi H. Mathematics in medicine: the 300-year legacy of iatro-mathematics. *Lancet*. 2013;382(9907): 1780.
3. Guerrini A. Archibald Pitcairne and Newtonian medicine. *Med Hist*. 1987;31:70–83.
4. Iatro-mathematics. *Lancet* 1920;196:610–11.
5. Putting Watson to Work: Watson in healthcare. IBM. Retrieved 11 Nov 2013.
6. IBM Watson helps fight cancer with evidence-based diagnosis and treatment suggestions. IBM. Retrieved 12 Nov 2013.
7. Saxena M. IBM Watson progress and 2013 Roadmap (Slide 7). IBM; 2013. Retrieved 12 Nov 2013.
8. Wakeman N. IBM's Watson heads to medical school. Washington Technology; 2011. Retrieved 19 Feb 2011.
9. Upbin B. IBM's Watson gets its first piece of business in healthcare. Forbes; 2013, February 8.
10. Miliard M. Watson heads to medical school: Cleveland Clinic, IBM Send Supercomputer to College. *Healthcare IT News*; 2012, October 30. Retrieved 11 Nov 2013.
11. Ghosh S. Google is consolidating DeepMind's healthcare AI business under its new Google Health unit. *Business Insider*. Retrieved 30 Jan 2020.
12. Baraniuk C. Google's DeepMind to peek at NHS eye scans for disease analysis. BBC; 2016, 6 July. Retrieved 6 July 2016.
13. Baraniuk C. Google DeepMind targets NHS head and neck cancer treatment. BBC; 2016, August 16. Retrieved 5 Sept 2016.
14. Marr B. Accessed 8 Feb 2021. <https://www.bernardmarr.com/default.asp?contentID=1373>
15. <https://medium.com/better-programming/pythons-advantages-and-disadvantages-summarized-212b5fdf8883>. Accessed 8 Feb 2021.
16. The 6th DOMO Report. Domo.com, 2018.
17. Andrae A, Edler T. On global electricity usage of communication technology: trends to 2030. *Challenges*. 2015;6:117–57.
18. Illustration by Nilay Nishit, Birla Institute of Technology, Mesra, India, May 2019.
19. Illustration by Federica Aresu, KTH, and Niklas Lidströmer, Karolinska Institute, Stockholm, Sweden, February 2021.
20. <https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm>, open-source image.
21. Illustration by Venkata Jagannath, TIBCO Spotfire, <http://community.tibco.com>. Accessed 12 Feb 2021, release as free license on Wikipedia.
22. <http://Machinelearningmastery.com>. Accessed 11 Feb 2021.
23. Oliver Carloni, SemSpirit.com, Research engines in artificial intelligence. His website presents a profound and comprehensive guidance and illustration of most of the useful calculations for AI.
24. Poggio T, Liao Q, Theory I. Deep networks and the curse of dimensionality. *Bull Polish Acad Sci Tech Sci*. 2018;66(6):761–73.
25. Rosenblatt F. The perceptron, a perceiving and recognizing automaton Project Para. Cornell Aeronautical Laboratory; 1957.
26. Image by Kiyoshi Kawaguchi, The University of Texas at El Paso College of Engineering Electrical & Computer Engineering, utep.edu.
27. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Neural Information Processing Systems (NIPS)*; 2012.
28. Sagar R. OpenAI releases GPT-3, the largest model so far. *Analytics India Magazine*. 2020, June 3. Retrieved 31 July 2020.
29. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Pfau D, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. Language models are few-shot learners. 2020;arXiv:2005.14165.
30. Document-feature matrix: Tutorials for quanteda. <http://tutorials.quanteda.io>. Accessed 11 Feb 2021.



Applying Principles from Medicine Back to Artificial Intelligence

2

Howard Schneider

Contents

Introduction	22
Concepts from the Medical Sciences Being Applied <i>Back to the Field of Artificial Intelligence</i> : Neural Networks	23
Concepts from the Medical Sciences Being Applied <i>Back to the Field of Artificial Intelligence</i> : Cognitive Architectures	24
Concepts from the Medical Sciences Being Applied <i>Back to the Field of Artificial Intelligence</i> : The Causal Cognitive Architecture	26
Discussion	31
References	32

Abstract

As artificial intelligence (AI) advances emerge, they are often incorporated into the field of medicine. However, the reverse of this process is an important one also—many of the breakthroughs in the field of AI are due to knowledge and inspiration *from* biology, psychology, and encompassing medical sciences. The artificial neuron, one of the first achievements in AI, is discussed. Knowledge about the visual cortex

inspiring the creation of the neocognitron leading to the convolutional neural network, leading to the deep learning revolution in AI, is discussed. Cognitive architectures, which intentionally incorporate knowledge from the medical sciences, including biology and psychology, into a form that can add to the field of AI, are reviewed. While deep learning models can recognize patterns and play games at a human-like level, they are poor at causal understanding and solutions, especially if there are limited numbers of training examples. The causal cognitive architecture is reviewed. This architecture stores information in the form of navigation maps, which it can update and link/retrieve to/from other navigation maps, and produces precausal behavior. If intermediate maps from its navigation center are fed back, stored, and operated on again, then full causal

Supplementary Information: The online version of this chapter (https://doi.org/10.1007/978-3-030-64573-1_289) contains supplementary material, which is available to authorized users.

H. Schneider (✉)
Sheppard Clinic North, Toronto, ON, Canada
e-mail: hschneidermd@alum.mit.edu

behavior can emerge. The architecture provides hypotheses and predictions about the emergence of schizophrenia in humans. The ability to navigate a world of concepts, explainability, and analogies also emerge from this architecture. Inspiration from the medical sciences can continue to provide AI with future breakthroughs.

Keywords

Artificial intelligence · Causal · Causality · Cognitive Architecture · Deep Learning · Explainability · Machine Learning · Perceptron · Psychosis · Schizophrenia

Introduction

As advances in artificial intelligence have developed over the last half century, they have been applied to the field of medicine with the hopes of automating many processes in medicine, resulting in better research towards an understanding of diseases and their treatments and a lower cost, higher quality health care service to patients. Consider, for example, the development of expert systems and deep learning and their application to medicine.

An expert system obtains knowledge, essentially if-then rules, from a human expert about some field and contains an inference engine that logically uses the if-then rules to answer questions or solve problems posed to it. Buchanan and colleague's DENDRAL expert system, which helped to identify an unknown organic chemical, was created in the late 1960s [1]. By the early 1970s, a medical application of an expert system, MYCIN, had been developed [2]. In response to answers to questions it asked, MYCIN could identify bacteria causing severe infections and recommend specific antibiotic treatments [3]. MYCIN had several hundred rules, used a straightforward inference engine, and produced reasonable responses for its very narrow range of expertise. However, it never was used in a practical clinical sense. Expert systems such as MYCIN raised hopes for the practical use of artificial intelligence in all aspects of science, technology, and

commerce. By the 1980s, expert systems had become a major focus of the field of artificial intelligence, and international commerce and competition in the field had started. However, by the early 1990s, it had become apparent that while it was possible to build interesting demonstration applications with expert systems, they largely failed to produce useful results for the complex tasks they had been promised for, medical applications included. The early 1990s period is often called the "AI winter," with analogy to the term "nuclear winter," in which the expert system companies failed, and institutional research funding was cut back in the field [4].

Deep learning generally refers to machine learning using many layers of artificial neural networks (ANNs); i.e., there are middle layer(s) of artificial neural networks between the input and output layers. The theory and technology behind various machine learning approaches, including ANNs with multiple hidden middle layers, had started to greatly improve in the mid-2000s. In 2012, work by Krizhevsky, Sutskever, and Hinton using deep learning won a computer vision competition by a large margin over older methods [5]. In the ImageNet contest, a computer-based system needed to classify as accurately as possible over a million different images into some thousand different classes. Hinton and colleagues used a deep convolutional neural network. In such a neural network, there are multiple layers of artificial neurons connected with layers where the artificial neurons act as convolutional layers where such layers extract features from the previous layers, preserving the spatial relationships but essentially mapping into a small-size receptive field and extracting features as such. This achievement of Hinton and colleagues is regarded as an approximate start of what is called the "deep learning revolution" and propelled the utilization of deep learning into many domains, including, of course, just about every branch of medicine. For example, consider the field of schizophrenia research and patient care. A review by Veronese and colleagues in 2013 gives an overview of machine learning approaches in schizophrenia but describes little of deep learning [6]. However, within a few years, deep learning was being widely used in the field. In 2016, Kim and

colleagues noted that deep neural networks (DNNs) with multiple hidden layers were performing much better in classification tasks compared to support vector machines (SVMs) and earlier AI models. Kim and colleagues used a DNN classifier of resting-state functional magnetic resonance imaging to diagnose schizophrenia patients from healthy controls and showed better results than the same analysis performed with an SVM classifier [7].

As artificial intelligence enhancements and technologies emerged, they eventually became incorporated into the field of medicine, reaching the point where there is now a scientific discipline of “artificial intelligence in medicine.” The journal *Artificial Intelligence in Medicine* defines the field as “the scientific discipline pertaining to research studies, projects and applications that aim at supporting decision-based medical tasks through knowledge-and/or data-intensive computer-based solutions that ultimately support and improve the performance of a human care provider” [8]. However, in this chapter, the reverse idea is considered. Rather than advances in the field of artificial intelligence being applied to medicine, there is examination here of the principles and advances in the medical sciences, including biology and psychology, being applied *back* to the field of artificial intelligence.

In this chapter, some examples of concepts from the medical sciences being applied *back* to the field of artificial intelligence are reviewed (Video 1):

- (i) Neural networks
- (ii) Cognitive architectures
- (iii) Causal cognitive architectures

Concepts from the Medical Sciences Being Applied Back to the Field of Artificial Intelligence: Neural Networks

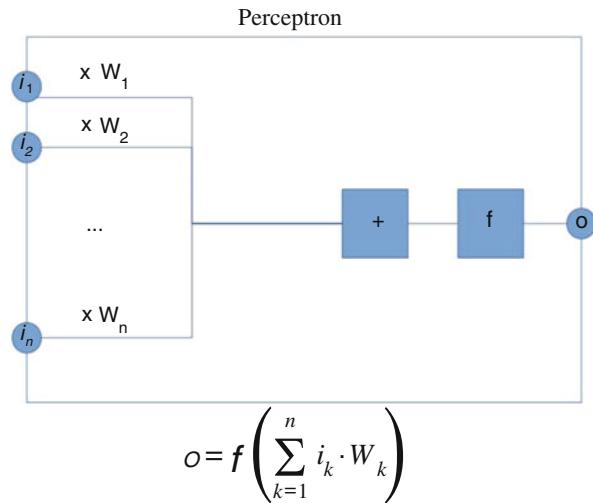
The basis of the “deep learning revolution” described above is the artificial neural network. As is obviously implied by the name, neural networks were inspired by biological neurons.

Neural networks were one of the earliest contributions to the field of AI. In 1943, Warren McCulloch and Walter Pitts took the basics of biological neuronal functioning and applied this knowledge to artificial neurons, which could be on or off and could activate other neurons [9]. They described a “logical calculus” for neural networks of their artificial neurons. A few years later, Donald Hebb described a method on how neurons could make connections with each other, i.e., “neurons that fire together wire together” [10], and an early learning mechanism arose that could be used with artificial neurons.

The artificial neuron, of course, does not simulate even a fraction of the complexity of an actual real neuron, but what it does do was sufficient for the emergence of neural networks. Rosenblatt’s “perceptrons” continued the use of these simplified neurons as functional elements of networks that could learn to classify inputs as belonging to a particular class [11]. Figure 1 shows a functional schematic of a perceptron. Inputs i_1 to i_n are multiplied by weights w , which are summed, and then passed through a function producing output o . Unfortunately, due to limitations in the way the perceptron networks were constructed (e.g., single-layer perceptron machine), it was shown that these simple neural networks could not function well enough to solve interesting problems [12]. Years later, more powerful and functional multilayer deep learning neural networks emerged, all with roots in the biological origins of the work of McCulloch and Pitts and which, as noted above, at the time of this writing are responsible for many of the modern applications of artificial intelligence, including applications in medicine.

Work by Hubel and Wiesel on the receptive fields of simple and complex cells in the visual cortex [13] inspired Fukushima’s neocognitron [14]. The neocognitron has simple (S) and complex (C) cells, with local features discerned by the simple cells and integrated into higher layers, and as such detects patterns. The neocognitron in turn inspired the convolutional neural network, which, as noted above, was used by Krizhevsky, Sutskever, and Hinton in their deep learning neural networks [5], which is often taken as the start of what is called the “deep learning revolution”

Fig. 1 Functional diagram of a perceptron, i – input, w – weight, f – function, o – output. (Creative Commons License BY-SA. Credit to Mat W.)



and resulted in the use of deep learning in many domains, including medicine.

Concepts from the Medical Sciences Being Applied Back to the Field of Artificial Intelligence: Cognitive Architectures

A cognitive architecture attempts to formally put a myriad of results from the medical sciences, including biology and psychology, into a model that gives a structure and functioning of the mind. As a result of the formal nature of the model, it can be implemented as a computer program, i.e., a computer program that exhibits varying degrees of intelligence. Concepts from the medical sciences, including biology and psychology, thus end up being applied back to the field of artificial intelligence.

ACT-R is a long-standing popular cognitive architecture [15]. ACT-R models the human mind as declarative knowledge (“chunks” which represent properties and are held in buffers) and procedural knowledge (“productions” which represent procedures to do various operations). There are two main types of long-term memory modules—declarative memory (i.e., facts, e.g., the pencil is brown) and procedural memory (i.e., productions, e.g., how to sharpen a pencil). Unlike typical databases in

computer science at the time which stored mainly facts, note that ACT-R’s memory is or can be substantially comprised of productions, i.e., procedures, as well. Just as humans have a short-term working memory, ACT-R also has a working memory. Depending on the internal state of ACT-R, which depends on the values in its various modules, including modules that interface with the real world, a best production will be selected and executed, and this in turn will change the internal state of the ACT-R, and then another production will be selected and executed, and so on. A cognitive architecture such as ACT-R is not just a collection of concepts but exists as a computer program, which can be run as well as modified. ACT-R has evolved over the years from 1973 to 2019 incorporating additional, modified, or newer theories of cognition, and also, its software implementation has changed [16].

The operation of a cognitive architecture such as ACT-R is inspired by the medical sciences. For example, in a recent ACT-R model, in its production system, the matching of productions is inspired by the striatum, the selection of a production is inspired by the pallidum, and the execution of a production is inspired by the thalamus [16]. In turn, the ACT-R contributes to the field of artificial intelligence, in which there are many diverse applications, e.g., human-robot interactions [17] or flying an airplane [18].

A variety of cognitive architectures exist. Samsonovich in 2010 and Kotseruba and colleagues in 2016 reviewed many of these different architectures [19, 20]. Ritter and colleagues organized cognitive architectures into five main groups [16]:

1. Cognitive architectures using advanced knowledge structures such as plans, with an emphasis on performance, e.g., JACK multiagent systems [21].
2. Symbolic architectures with emphasis on modeling human cognition, e.g., Soar cognitive architecture [22] (although some versions of Soar also have subsymbolic operations and would be thus classified in the hybrid group below)
3. Subsymbolic/connectionist architectures with information distributed across multiple nodes, e.g., Leabra cognitive architecture [23].
4. Hybrid architectures with both symbolic and subsymbolic components, with emphasis on modeling human cognition, e.g., ACT-R, since it has symbolic production and declarative memory components, but it also has subsymbolic operations in the operation of the architecture
5. Nongenerative architectures that do not produce behavior but are largely used as design tools to predict time and other requirements to accomplish some set of procedures.

Although O'Reilly and colleagues' work describes in particular the Leabra architecture, their work applies more broadly to the design of other cognitive architectures [23]. They note that cognitive architectures represent a complex way to do computational modeling, but the reason they are used is that modeling human cognition does not seem possible with a simpler generalized algorithm. O'Reilly and colleagues give a list of principles that they used to develop the Leabra architecture and are broad enough to be used for other architectures. Some of these principles are:

1. Balance the tradeoffs associated with different approaches, often using a compromised approach that may integrate multiple solutions.

2. Biology is important. The human brain is the only working truly successful cognitive system, so there is merit in trying to understand its details.
3. Occam's razor—regardless of the complexity required, it is best to use the simplest model that is sufficient for the modeling required.
4. Again, there is a need to balance tradeoffs between biological constraints, cognitive constraints, and computational constraints, depending on the model desired.
5. Experience-driven learning mechanisms are essential.
6. Microstructural mechanisms will affect the macrostructural mechanisms.
7. Changes in neural firing, i.e., activation, can occur more quickly than changes in synapses.
8. Meaning, in a connectionist system, is due to the patterns across the entire system, not individual neural messages.

An interesting cognitive architecture is the very hybridized OpenCog, which has the goal of creating a human-equivalent artificial general intelligence (AGI) [24–26]. OpenCog uses a graphical database that links to a myriad of different cognitive processes, including forward and backward chaining of essentially production rules, Bayesian inference (i.e., update probabilities as more information becomes available), an economic attention allocation mechanism where there are numerous attention values attached to various pieces of information, probabilistic logic networks (allow reasoning with uncertain information in the real world), meta-optimizing semantic evolutionary and probabilistic evolutionary mechanisms (where “evolutionary” mechanisms are not actual biological genetics but refer to the field of genetic programming, i.e., of applying fitness selection to large numbers of, for example, random-like programs and selecting and evolving the more fit programs), a natural language input and output processing system, and the use of emotions. While at this time of writing OpenCog has obviously not even come close to its goal of a human equivalent AGI, and while OpenCog contains a potpourri of cognitive concepts, many of these do stem originally from animal and human

minds and show the potential of applying principles from the medical sciences *back* to the field of artificial intelligence.

Concepts from the Medical Sciences Being Applied *Back* to the Field of Artificial Intelligence: The Causal Cognitive Architecture

As noted above, a cognitive architecture incorporates results from the medical sciences, including biology and psychology, into a system that models a mind, whether natural or artificial. This model is formal enough that it can be implemented as a computer program. Like other cognitive architectures, what is termed the “causal cognitive architecture” arises from observations in the medical sciences [27–32]. A simplified block diagram of an implementation of a causal cognitive architecture, the Causal Cognitive Architecture 1 (CCA1), is shown in Fig. 2 [32]. Some of the many observations and concepts from the medical sciences, including biology and psychology, that the causal cognitive architecture incorporates will be discussed first. Then its basic properties as well as its emergent properties will be considered.

In most of the animal world, in both invertebrates and vertebrates, there is the ability to move and the ability to navigate. In mammals, Hafting and colleagues [33] showed neural maps of the spatial environment in the dorsocaudal medial entorhinal cortex, with location activation of “grid cells” in these maps. Schafer and Schiller note that it is possible that these maps intended originally for navigation in the physical world can be used to navigate concepts [34]. Evolutionary precursors to the mammalian cortex go back to the earliest vertebrates [35]. Thus, in the design of the causal cognitive architecture, it was decided that information would be stored as navigation maps. Operations on these maps could be effected in a “navigation module” where maps would be modified, updated, stored adjacently, retrieved, linked to other maps, and so on. Except for some typical neural network-like hierarchical processing of input signals and output signals, all information would be stored and manipulated in map-like data

structures, and much of the higher processed information would pass through the navigation module [32].

The mammalian cortex appears to be made up of large numbers of repeating cortical minicolumns [36]. Also, there is a generative aspect in how the mind processes information. As a result, a basic pattern recognizing circuit, a Hopfield-like network (HLN), was chosen for many of the causal cognitive architecture’s circuits. HLNs can be dynamically reconfigured with other HLNs to extract what is described in the architecture as maximal “meaningfulness” from input vectors, where meaningfulness is the reciprocal of the Shannon entropy, favoring activation of the maximal number of HLNs further downstream [27, 31, 32].

Just as a child is not a tabula rasa, neither is the causal cognitive architecture. The instinctive primitives module shown in the architecture in Fig. 2 contains procedural vectors, which can be triggered by processed sensory inputs, by the navigation module, or by other modules of the architecture. If triggered, an instinctive primitive will be applied to the navigation map currently in the navigation module and can, for example, move along the map to trigger another instinctive or learned primitive or change the map or cause another linked map to be retrieved and so on. The instinctive primitives include physics primitives, mathematical primitives, psychology primitives, and planning primitives. Also, there are learned primitives that as the name implies are learned from experience. There is a letter “D” in many of the boxes in the architecture in Fig. 2. “D” stands for the internal developmental timer; i.e., as the CCA1 shown in Fig. 2 gains more experience, different levels of instinctual primitives or operations in these modules will be executed.

Examining Fig. 2, note that the sensory inputs are processed by hierarchies of input sensory vector association modules and then a binding module. Inputs are then sent to the navigation module. (They also go to the learned and instinctive primitives where they can trigger primitives to be executed, to the sequential/error correcting module which is useful for sequential data, to the autonomic modules, and to other modules.)

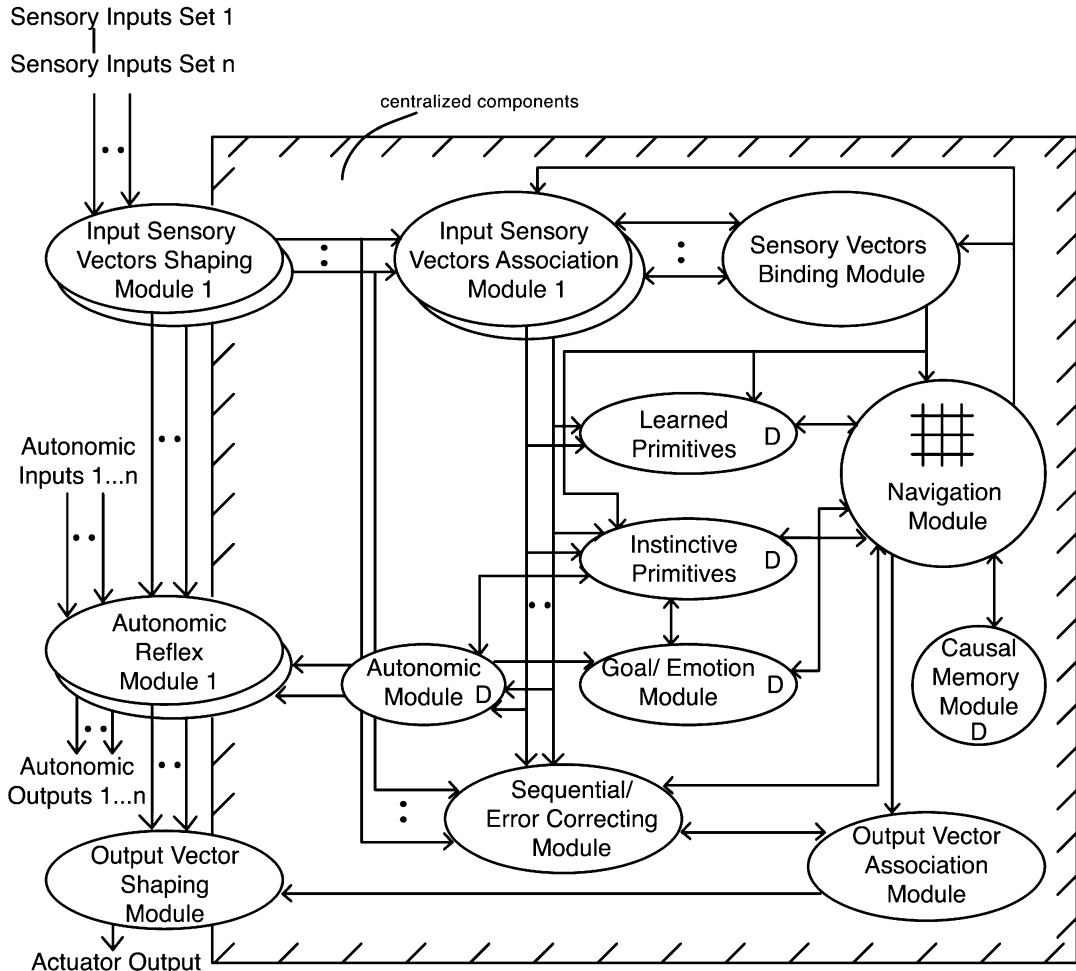


Fig. 2 Causal Cognitive Architecture 1 (CCA1) Not all connections are shown. D – internal developmental timer. (License – Original work for this publication)

Applying primitives, i.e., vectors that cause a modification of the current map in the navigation module, allows the input signal to be modified in a precausal fashion. The map already has some causality built into it; i.e., this goes to that, which goes to that and so on. The output of the navigation module is processed, in its most basic form, via the output vector association and shaping modules, into a movement of an actuator. As such, precausal behavior emerges from the Causal Cognitive Architecture 1 shown in Fig. 2 [32].

Deep learning can recognize patterns and use reinforcement learning to operate at human-like levels for certain skills such playing games [37, 38]. However, if deep learning needs to

causally make sense of a problem, especially if there are not large numbers of training examples, its performance is poor compared to a four-year-old child [39, 40]. In the causal cognitive architecture, instinctive and learned primitives applied against map-like memories or input signals represented in map-like structures will produce outputs that the maps lead to, which as noted above results in a precausal complex associative behavior. Many birds as well as mammals can exhibit precausal behavior, i.e., not a full understanding of cause and effect but a complex association of causes with outcomes that is more than just a simple reflex or association. However, it is unclear whether *any* nonhumans can effectively

understand cause and effect and can truly use full causality in making decisions.

Asian elephants have brains much larger than humans, but in experiments by Nissan [41] behavior was due to associative learning rather than causal learning or behavior. New Caledonia crows are frequently claimed to show causal behavior in solving physical problems. However, in research by Neilands and colleagues where the crows must drop a weighted object down a tube, the crows did not understand causally the idea of force, and there was little causal understanding overall [42]. Tayler and colleagues examined crows solving string pulling problems and came to the conclusion that there was little causal understanding but that the problems were solved instead by a perceptual-motor feedback loop [43]. If the behavior of primates, obviously closer in origin to humans, is considered, there still is not a clear indication that robust causal behavior occurs. For example, Visalberghi and Limongelli showed that while capuchin monkeys were able to use a stick to push a reward out of tube, if a hole and a trap are added to the tube, the monkeys showed poor understanding of cause-and-effect relations in not pushing the food into the trap [44]. This tube with a trap and food-reward-type problem is also difficult for chimpanzees. However, work by Seed [45] and colleagues showed that if the tube problem is simplified a bit, there appears to be some understanding by the chimpanzees about the functional properties of the problem. However, while the chimpanzees were able to solve this simplified problem, it still required dozens of trials [45].

While there remains controversy concerning the extent of causal behavior possible in non-humans, note that chimpanzees are the closest extant relatives to humans and yet are not capable of a robust understanding of cause and effect or causal behavior. Thus, a design requirement of the causal cognitive architecture was that there should be a simple mechanism to explain how the architecture can transition from the precausal behavior described above to a fully causal behavior [32].

Almost all the circuits of the causal cognitive architecture have feedback pathways. Feedback from downstream hierarchical levels as well as

lateral levels send back signals which will allow the CCA1 to better anticipate what the next input sensory vector should be and thus better recognize such input signals. In Fig. 2 showing the Causal Cognitive Architecture 1, if the feedback pathways from the navigation module to the input sensory various modules are enhanced, then full intermediate navigation maps from the navigation module can be stored temporarily in the input sensory modules. In the next input cycle, rather than considering actual sensory input signals, the intermediate results can be fed back to the navigation module and operated on again by the instinctive and learned primitives [32].

The ability of the causal cognitive architecture to temporarily store intermediate values of the navigation module and then process these map-like structures and values again, and again as many times as necessary, results in the ability of the causal cognitive architecture to have robust causal behavior [31, 32]. Note that the transition from precausal to full causal ability merely requires enhanced feedback pathways and a slight change in how input signals are processed, in keeping with the requirement above that the architecture be able to provide a mechanism for a simple transition from precausal to causal abilities.

In a simulation of the Causal Cognitive Architecture 1 (CCA1) controlling a search and rescue robot in a rain forest searching for a lost hiker [32], in precausal mode, if the robot comes to a fast-moving, noisy river, it will cross the river since it is able to cross rivers. As sensory inputs are processed in the navigation module, intuitive and learned primitives are activated and will not prevent the robot from entering the river. However, in this rain forest, the fast-moving, noisy river is the start of a waterfall, and unfortunately the robot is swept over the edge of the waterfall and is damaged. If the robot is repaired and goes out into the rain forest and senses another fast-moving, noisy river, then by associative mechanisms, i.e., via the learned primitives (as well as the goal/emotion module), it will avoid the river.

In the same situation in full causal mode [32], “fast moving” + “water” will cause an intuitive physics primitive to cause the navigation module

to produce the navigation map “push” + “water” (where there are not actually discrete symbols but instead a map of the robot moving in water). The intermediate navigation map “push” + “water” is fed back to the sensory circuits, and in the next cycle processing input signals, the “push” + “water” navigation map loads back in the navigation module and is operated on by the intuitive and learned primitives. The result is pulling up or creating a new map with the robot underneath water. This intermediate result is fed back and then used as the input signal in the next cycle. An intuitive primitive is triggered by this map of the robot underneath the water, with an output from the navigation center to not go in this direction. As a result, in full causal mode, the search and rescue robot does not enter this fast-moving, noisy river and become damaged, even though it has never seen a waterfall or a noisy fast-moving river before or trained on such examples.

An emergent property of the causal cognitive architecture is the ability to not only handle navigation in the physical world but also use the map-like data structures in the navigation center to hold and link to more abstract concepts and to navigate this world of concepts. Causal operations can be performed on these concepts as they are modified and stored and cause retrieval of other concepts [32].

The navigation module shown in Fig. 2 stores the many navigation maps it constructs in the causal memory module. When a similar situation occurs again in the future, then similar navigation maps are triggered in the causal memory module and fed into the navigation module, thus providing a learned representation of which actions occurred in the past and which can be used for the new situation, with modifications as needed. An emergent property of this is explainability—the sequence of navigation maps used in an action or decision gives a very reasonable explanation of why a particular answer was given for a particular problem. (Note that the subsymbolic aspects of the causal cognitive architecture, such as the initial steps in processing an input signal through a hierarchy of Hopfield-like networks or equivalent, are not fully captured by the navigation maps.) Gilpin and colleagues [46] describe the property

of explainability as a model being able to give the reasons for its behavior. The lack of explainability in conventional deep learning models is an important concern when using such models for critical applications, including those of AI in medicine.

Another emergent property of this architecture is the ability to generate and handle analogies [32]. For example, consider asking the almost philosophical question to the search and rescue robot in the example above after returning from its rescue mission of whether it wants to spend time with either person A or person B, who are both similar but person B is more smiley and more outgoing, i.e., very noisy. Person A and person B would be put onto a temporary map in the CCA1’s navigation module (Fig. 2). The Causal Cognitive Architecture 1 has to essentially decide whether to navigate to person A or person B. Its instinctive primitives, which include psychology primitives, favor smiling people. However, person B is also noisy, and this results in the navigation center pulling up a previous navigation map—the river was noisy, and thus the link on the map with person B now ends up pointing to a dangerous situation. Intermediate results would be fed back to the sensory input stages and processed again, temporary maps would be switched back again, and the noisy person B is now associated with possible danger. Thus, there is a navigation output to navigate to person A. Note that without any elaborate, specialized algorithms and without any special central controlling stored program, other than the inherent and relatively simple operational cycles of the architecture, the CCA1 has made what would seem like a cognitively advanced decision. Analogies readily emerge from the causal cognitive architecture.

The causal cognitive architecture was not intended to model disease, and indeed no pathological predispositions were intentionally designed into its architecture. However, an interesting emergent property of the architecture is that when the mode switches from precausal to fully causal (i.e., send back intermediate results from the navigation module to the sensory stages so they can be operated on again in the next sensory input cycle), psychotic behavior may emerge.

Any number of varied imperfections or combinations of them can result in psychotic behavior (but in precausal mode usually would not have had catastrophic effects as such) [29, 31, 32]. The intermediate navigation maps that the navigation module stores temporarily in the input sensory vector circuits, under a faulty operation, can be interpreted as a real input sensory signal, i.e., effectively hallucination-like and then delusional-like responses when instinctive and learned primitives are applied to it in subsequent operations, and then cognitive dysfunction of the system. Hallucinations, delusions, and cognitive dysfunction are typical of psychosis. Note that it is not just one defect that can cause this, but many different defects in various circuits.

Since the causal cognitive architecture is inspired by biology (although *not* pathology), then if it predicts that psychosis easily emerges, why is it that only approximately 1% of the human population has schizophrenia and not a larger percentage of the population? Actually, as van Os and colleagues show in their research [47], more than 10% of the population (a large figure from a population point of view) will experience some sort of psychotic-like symptoms—there are many causes why humans may experience psychotic-like symptoms, as the causal cognitive architecture predicts. Work by Anttila and colleagues [48] looking at the genomes of 265,218 psychiatry patients and 784,643 controls found considerable genetic overlap between what should be very different formal psychiatric disorders, including schizophrenia, again in keeping with the causal cognitive architecture predicting that psychosis does not emerge from a single or small group of genes.

In the causal cognitive architecture, in the non-causal mode, i.e., in the precausal mode, the psychotic-like behavior does not happen as easily. Note that humans readily develop psychosis (i.e., the 10% figure above) but humans do have robust causal abilities. While there is still controversy concerning the extent of causal behavior possible in nonhumans, from all the evidence available, as discussed above, nonhumans do not have robust causal behavior. Should psychosis readily arise in nonhumans? The causal cognitive

architecture would predict no. In fact, in just about all other mammals, psychosis is rare, and in psychopharmacological research settings, large efforts are needed to induce at best unreliable models of schizophrenia in research animals [49].

Causal aspects of the causal cognitive architecture are inspired by the abilities of human working memory. The causal cognitive architecture would predict that in asymptomatic relatives of patients with psychosis, there are also defects in the causal process and the working memory, even though perhaps in that person the cumulative effects of the defects do not result in psychosis. In fact, decreased working memory abilities are found not only in patients with schizophrenia but also in unaffected relatives [50].

It is interesting to look at the transition from nonhumans, where the closest extant relatives of humans, chimpanzees, do not have full causal abilities nor readily develop psychosis, and it would be assumed that the last common ancestors of both modern humans and chimpanzees did not either. The “schizophrenia paradox” is that schizophrenia reduces an individual’s ability to reproduce and should be naturally eliminated from the population yet continues to be found throughout the world at a relatively high prevalence approaching 1%. There are all sorts of rationalizations for this [51–54]. However, the causal cognitive architecture rejects the schizophrenia paradox and these explanations. Instead, it predicts that the transition to a brain architecture that allows causality is also vulnerable to many different defects allowing psychosis to emerge. Thus, there should be many different genetic alleles that prior to the transition to full causality were not harmful, that now would allow psychosis to emerge more easily, and that removing a myriad of different genetic characteristics from a population will not occur as easily as removing a single faulty allele [29, 31, 32].

Work by Liu and colleagues [55] published in 2019 compared single nucleotide polymorphisms (SNPs) associated with schizophrenia in modern humans with those present in the recovered genomes of extinct archaic Denisovan and Neanderthal humans (thought to have split from the modern human ancestors approximately 440,000

to 270,000 years ago [56]). Liu and colleagues showed that the risk alleles for schizophrenia appear to have been gradually removed from the modern human genome due to the negative selection pressure, as the causal cognitive architecture predicted.

Discussion

As artificial intelligence advancements and technologies emerge, they are often incorporated into the field of medicine. However, it is noted that the reverse of this process is an important one also. Many of the large improvements and breakthroughs in the field of artificial intelligence are in fact due to knowledge and inspiration *from* the medical sciences, including biology and psychology. A number of examples were discussed above.

The basis of neural networks, the artificial neuron, was one of the first achievements in the field of artificial intelligence. It was derived from knowledge in the biological and medical fields. Similarly, scientific knowledge about the visual cortex inspired the creation of the neocognitron. In turn, the neocognitron led to the convolutional neural network, which in turn led the start of what is sometimes called the deep learning revolution. Deep learning neural networks have at this time been applied to almost every field of academia, industry, and, of course, medicine.

Cognitive architectures intentionally attempt to take knowledge from the medical sciences, including biology and psychology, and incorporate this knowledge into a formal structure, which then can be implemented as a computer program. A variety of cognitive architectures were discussed above. While many cognitive architectures make only small contributions to the field of AI but may be more useful to provide insight into the operation of the human mind, some are intended as bona fide contributions to artificial intelligence. The OpenCog cognitive architecture, for example, has the goal of becoming a human equivalent artificial general intelligence, although admittedly at the time of writing, it is far from this achievement.

As was noted above, deep learning can recognize patterns and use reinforcement learning to play games at a human-like level [37, 38]. However, if a deep-learning-based system needs causal understanding to solve a problem, especially if there are not large numbers of training examples, its performance is poor compared to a 4-year-old child [39, 40]. For many situations, it simply is not possible or practical to have extensive training on every type of example that could arise. This certainly happens in the medical applications of AI as well as in almost all other fields as well. Work by Amodei and Hernandez [57] note that the famous Moore's law has a 2-year doubling period (i.e., the number of transistors on integrated circuits fabricated by manufacturers doubles approximately every 2 years), which means from 2012 to 2018 there should have been an approximately eight-fold improvement in the power of computer hardware. A 800% increase is, of course, significant and impressive. However, what really happened in the training of deep learning networks from 2012 to 2018 was that a 300,000-fold increase (i.e., a 30 million % increase) in hardware computational power was used to train deep learning networks. This number has only increased in the years since 2018. These exponential increases in computing power are not sustainable, nor is the ever-increasing need for more training examples. The resources and costs simply become prohibitive, despite the improvements by hardware manufacturers and despite the improvements in deep learning training algorithms. Thus, deep learning neural networks may start approaching pragmatic limits in their ability to do more complex tasks at a human level, certainly those requiring causal decisions and limited training examples.

Others in the field of AI have started to recognize the problems that current implementations of deep learning face and the need to consider alternative approaches. Discussed above, for example, was the interesting cognitive architecture OpenCog, which has the goal of creating a human-equivalent artificial general intelligence [24–26]. Graves and colleagues [58] discuss using a neural network that can read and write to an external memory, i.e., a hybrid system. Huyck [59] describes work on a neuromorphic-like

cognitive architecture with a fast implicit subconscious system and a slow explicit conscious system. Epstein [60] discusses cognitive modeling of spatial navigation. Lake and colleagues [61] discuss building causal models of the world and discuss intuitive physics and psychology present in infants. Hawkins and colleagues [62] and others, such as Schafer and Schiller [34], discuss how abstract concepts can be represented in a spatial framework. Laird and Mohan [63] discuss using architectures that contain more innate learning mechanisms. Taatgen [64] discusses using multiple levels of abstraction for learning. In a recent paper discussing deep learning for higher-level cognition [65], Goyal and Bengio write: “Have the main principles required for deep learning to achieve human-level performance been discovered, with the main remaining obstacle being to scale up? We argue that having larger and more diverse datasets is important but insufficient without good architectural inductive biases.”

While the causal cognitive architecture discussed above is largely at an experimental level, it is a technology that is essentially derived from the medical sciences, including biology and psychology [32]. As noted above, the causal cognitive architecture creates navigation maps from input sensory vectors and produces a navigation output, i.e., related to movement. A precausal behavior can result from this architecture. There is no understanding of cause and effect, but the maps navigate from problem to solution that may seem to be related to this or that cause. However, intermediate solutions to a problem, i.e., the maps created in the navigation module, can be fed back along existing but enlarged feedback pathways and stored as an input sensory value, and in the next input cycle processed again, and repeatedly so as needed. Maps of intermediate parts of the problem solution can be created and stored, new maps made, previous maps retrieved, and so on. In this fashion, true causal behavior can result with an understanding of cause and effect.

The causal cognitive architecture, inspired and derived by the medical sciences, including biology and psychology, offers features, albeit in toy

models tested to date [27–32], useful for the field of AI. It allows problems to be mapped and processed causally. It stores the intermediate maps it creates in solving a problem or performing a behavior—retrieving and replaying the maps allow explainability about a decision the architecture made. While solving physical navigation problems is obvious, the causal cognitive architecture can similarly navigate a world of concepts. Another emergent property of the architecture is the ability to handle analogies automatically.

As shown above, many of the large improvements and breakthroughs in the field of artificial intelligence are in fact due to knowledge and inspiration *from* the medical sciences. An appreciation for the origin of the technologies in AI can help with their application in general as well as to the medical fields.

References

- Buchanan BG, Sutherland GL, Feigenbaum EA. Heuristic DENDRAL: a program for generating explanatory hypotheses in organic chemistry. In: Meltzer B, Michie D, Swann M, editors. Machine Intelligence 4 – Proceedings of the Fourth Annual Machine Intelligence Workshop. Edinburgh: Edinburgh University Press; 1969.
- Buchanan BG, Shortliffe EH. Rule based expert systems: The MYCIN experiments of the Stanford Heuristic Programming Project. Reading: Addison-Wesley; 1984.
- Perry CA. Knowledge bases in medicine: a review. Bull Med Libr Assoc. 1990;78(3):271–82. PMID: 2203499
- Russell S, Norvig P. The history of artificial intelligence. In: Artificial intelligence: a modern approach. 4th ed. Hoboken: Pearson; 2021. p. 17–27.
- Krizhevsky, A, Sutskever, I, Hinton, GE. ImageNet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems – Volume 1 (NIPS’12). Curran Associates: Red Hook, NY, USA, 1097–1105; 2012.
- Veronese E, Castellani U, Peruzzo D, Bellani M, Brambilla P. Machine learning approaches: from theory to application in schizophrenia. Comput Math Methods Med. 2013;2013:867924. <https://doi.org/10.1155/2013/867924>.
- Kim J, Calhoun VD, Shim E, Lee JH. Deep neural network with weight sparsity control and pre-training

- extracts hierarchical features and enhances classification performance: evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. *NeuroImage*. 2016;124(Pt A):127–46. <https://doi.org/10.1016/j.neuroimage.2015.05.018>.
- 8. Sciedirect.com/journal/artificial-intelligence-in-medicine. About the journal – aims and scope. 2020. [cited 2020 Dec 8]. Available from: <https://www.sciencedirect.com/journal/artificial-intelligence-in-medicine>
 - 9. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys*. 1943;5:117–37.
 - 10. Hebb DO. The organization of behavior. New York: Wiley; 1949.
 - 11. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev*. 1958;65(6):386–408. <https://doi.org/10.1037/h0042519>.
 - 12. Minsky ML, Papert SA. Perceptrons. Cambridge, MA: MIT Press; 1969.
 - 13. Hubel DH, Wiesel TN. Receptive fields of single neurones in the cat's striate cortex. *J Physiol* Oct. 1959;148(3):574–91. <https://doi.org/10.1111/jphysiol.1959.sp006308>.
 - 14. Fukushima K. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol Cybern*. 1980;36: 193–202. <https://doi.org/10.1007/BF00344251>.
 - 15. Anderson JR. The architecture of cognition. Cambridge, MA: Harvard University Press; 1983.
 - 16. Ritter FE, Tehranchi F, Oury JD. ACT-R: a cognitive architecture for modeling cognition. *Wiley Interdiscip Rev Cogn Sci*. 2019;10(3):e1488. <https://doi.org/10.1002/wcs.1488>.
 - 17. Trafton JG, Hiatt LM, Harrison AM, Tamborello FP, Khemlani SS, Schultz AC. ACT-R/E: an embodied cognitive architecture for human-robot interaction. *J Hum Robot Interact*. 2013;2:30–55. <https://doi.org/10.5898/JHRI.2.1.Trafton>.
 - 18. Schoppek W, Boehm-Davis DA. Opportunities and challenges of modelling user behavior in complex real world tasks. *MMI Interaktiv*. 2004;7:47–60.
 - 19. Samsonovich A. Toward a unified catalog of implemented cognitive architectures. In: Biologically inspired cognitive architectures; 2010. p. 195–244. <https://doi.org/10.3233/978-1-60750-660-7-195>.
 - 20. Kotseruba, I, Gonzalez, O, Tsotsos, JK. A review of 40 years of cognitive architecture research. arXiv: 1610.08062v3 [cs.ai]; 2016.
 - 21. Busetta P, Howden N, Rönnquist R, Hodgson A. Structuring BDI agents in functional clusters. In: Jennings NR, Lespérance Y, editors. Intelligent agents VI. Agent theories, architectures, and languages. ATAL 1999. Lecture notes in computer science, Vol. 1757. Berlin: Springer; 2000. https://doi.org/10.1007/10719619_21.
 - 22. Laird JE. The soar cognitive architecture. Cambridge, MA: MIT Press; 2012.
 - 23. O'Reilly RC, Hazy TE, Herd SA. The Leabra cognitive architecture: how to play 20 principles with nature and win! In: Chipman SEF, editor. The Oxford handbook of cognitive science. New York: Oxford University Press; 2017. p. 91–115.
 - 24. Hart D, Goertzel B. OpenCog: a software framework for integrative artificial general intelligence. In: Wang, et al., editors. Proceedings of the first AGI conference. Amsterdam, Netherlands: IOS Press; 2008. p. 468–72.
 - 25. Goertzel, B, Duong, D. OpenCog NS: a deeply-interactive hybrid neural-symbolic cognitive architecture designed for global/local memory synergy. AAAI Fall Symposium: Biologically Inspired Cognitive Architectures; 2009.
 - 26. Goertzel B, Pennachin C, Geisweiller N. Engineering general intelligence, Part 2: the CogPrime architecture for integrative, embodied AGI. Paris: Atlantis Press; 2014.
 - 27. Schneider H. Meaningful-based cognitive architecture. In: Samsonovich AV, editor. Biologically inspired cognitive architectures BICA 2018. Procedia computer science, vol. 145; 2018. p. 471–80. <https://doi.org/10.1016/j.procs.2018.11.109>.
 - 28. Schneider H. Subsymbolic versus symbolic data flow in the meaningful-based cognitive architecture. In: Samsonovich AV, editor. Biologically inspired cognitive architectures BICA 2019, Advances in intelligent systems and computing, vol. 948; 2020. p. 465–74. https://doi.org/10.1007/978-3-030-25719-4_61.
 - 29. Schneider H. Schizophrenia and the future of artificial intelligence. In: Samsonovich AV, editor. Biologically inspired cognitive architectures 2019, Advances in intelligent systems and computing, vol. 948; 2020. p. 475–84. https://doi.org/10.1007/978-3-030-25719-4_62.
 - 30. Schneider H. Emergence of belief systems and the future of artificial intelligence. In: Samsonovich AV, editor. Biologically inspired cognitive architectures BICA 2019, Advances in intelligent systems and computing, vol. 948; 2020. p. 485–94. https://doi.org/10.1007/978-3-030-25719-4_63.
 - 31. Schneider H. The meaningful-based cognitive architecture model of schizophrenia. *Cogn Syst Res*. 2020;5(9):73–90. <https://doi.org/10.1016/j.cogsys.2019.09.019>.
 - 32. Schneider H. Causal cognitive architecture 1: integration of connectionist elements into a navigation-based framework. *Cogn Syst Res*. 2021;66:67–81. <https://doi.org/10.1016/j.cogsys.2020.10.021>.
 - 33. Hafting T, Fyhn M, Molden S, Moser MB, Moser EI. Microstructure of a spatial map in the entorhinal cortex. *Nature*. 2005;436(7052):801–6. <https://doi.org/10.1038/nature03721>.

34. Schafer M, Schiller D. Navigating social space. *Neuron*. 2018;100(2):476–89. <https://doi.org/10.1016/j.neuron.2018.10.006>.
35. Suryanarayana SM, Robertson B, Wallén P, et al. The lamprey pallium provides a blueprint of the mammalian layered cortex. *Curr Biol*. 2017;27(21):3264–77. <https://doi.org/10.1016/j.cub.2017.09.034>.
36. Buxhoeveden DP, Casanova MF. The minicolumn hypothesis in neuroscience. *Brain*. 2002;125 (Pt 5):935–51. <https://doi.org/10.1093/brain/awf110>.
37. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, MA: MIT Press; 2016.
38. Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. *Nature*. 2015;518:529–33. <https://doi.org/10.1038/nature14236>.
39. Waismeyer A, Meltzoff AN, Gopnik A. Causal learning from probabilistic events in 24-month-olds: an action measure. *Dev Sci*. 2015;18(1):175–82. <https://doi.org/10.1111/desc.12208>.
40. Ullman S. Using neuroscience to develop artificial intelligence. *Science*. 2019;363(6428):692–3. <https://doi.org/10.1126/science.aau6595>.
41. Nissan M. Do Asian elephants (*Elephas maximus*) apply causal reasoning to tool-use tasks? *J Exp Psychol Anim Behav Process*. 2006;32(1):91–6. <https://doi.org/10.1037/0097-7403.32.1.91>.
42. Neilands PD, Jelbert SA, Breen AJ, Schiestl M, Taylor AH. How insightful is ‘insight’? New Caledonian Crows do not attend to object weight during spontaneous stone dropping. *PLoS One*. 2016;11(12):e0167419. <https://doi.org/10.1371/journal.pone.0167419>.
43. Taylor AH, Knaebe B, Gray RD. An end to insight? New Caledonian crows can spontaneously solve problems without planning their actions. *Proc Biol Sci*. 2012;279(1749):4977–81. <https://doi.org/10.1098/rspb.2012.1998>.
44. Visalberghi E, Limongelli L. Lack of comprehension of cause-effect relations in tool-using capuchin monkeys (*Cebus apella*). *J Comp Psychol*. 1994;108(1):15–22. <https://doi.org/10.1037/0735-7036.108.1.15>.
45. Seed AM, Call J, Emery NJ, Clayton NS. Chimpanzees solve the trap problem when the confound of tool-use is removed. *J Exp Psychol Anim Behav Process*. 2009;35(1):23–34. <https://doi.org/10.1037/a0012925>.
46. Gilpin, LH, Bau, D, Yuan, BZ, Bajwa, A, Specter, M, Kagal L. Explaining explanations: an overview of interpretability of machine learning. 2018 IEEE 5th international conference on data science and advanced analytics (DSAA), Turin, Italy, 2018;80–89. <https://doi.org/10.1109/DSAA.2018.00018>.
47. van Os J, Hanssen M, Bijl RV, et al. Prevalence of psychotic disorder and community level psychotic symptoms: an urban-rural comparison. *Arch Gen Psychiatry*. 2001;58(7):663–8.
48. Brainstorm Consortium, Anttila V, Bulik-Sullivan B, Finucane HK, Walters RK, Bras J, Duncan L, Escott-Price V, et al. Analysis of shared heritability in common disorders of the brain. *Science*. 2018;360(6395):eaap8757. <https://doi.org/10.1126/science.aap8757>.
49. Jones CA, Watson DJ, Fone KC. Animal models of schizophrenia. *Br J Pharmacol*. 2011;164(4):1162–94. <https://doi.org/10.1111/j.1476-5381.2011.01386.x>.
50. Zhang R, Picchioni M, Allen P, Toulopoulou T. Working memory in unaffected relatives of patients with schizophrenia: a meta-analysis of functional magnetic resonance imaging studies. *Schizophr Bull*. 2016;42(4):1068–77. <https://doi.org/10.1093/schbul/sbv221>.
51. Pearlson GD, Folley BS. Schizophrenia, psychiatric genetics, and Darwinian psychiatry: an evolutionary framework. *Schizophr Bull*. 2008;34(4):722–33. <https://doi.org/10.1093/schbul/sbm130>.
52. Benítez-Burraco A, Di Pietro L, Barba M, Lattanzi W. Schizophrenia and human self-domestication: an evolutionary linguistics approach. *Brain Behav Evol*. 2017;89(3):162–84. <https://doi.org/10.1159/000468506>.
53. Polimeni J, Reiss JP. Evolutionary perspectives on schizophrenia. *Can J Psychiatr*. 2003;48(1):34–9. <https://doi.org/10.1177/070674370304800107>.
54. Crow TJ. Schizophrenia as the price that *homo sapiens* pays for language: a resolution of the central paradox in the origin of the species. *Brain Res Brain Res Rev*. 2000;31(2–3):118–29. [https://doi.org/10.1016/s0165-0173\(99\)00029-6](https://doi.org/10.1016/s0165-0173(99)00029-6).
55. Liu C, Everall I, Pantelis C, Bousman C. Interrogating the evolutionary paradox of schizophrenia: a novel framework and evidence supporting recent negative selection of schizophrenia risk alleles. *Front Genet*. 2019;10:389. <https://doi.org/10.3389/fgene.2019.00389>.
56. Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PL, Maricic T, Good JM, Marques-Bonet T, Alkan C, Fu Q, Mallick S, Li H, Meyer M, Eichler EE, Stoneking M, Richards M, Talamo S, Shunkov MV, Derevianko AP, Hublin JJ, Kelso J, Slatkin M, Pääbo S. Genetic history of an archaic hominin group from Denisova cave in Siberia. *Nature*. 2010;468(7327):1053–60. <https://doi.org/10.1038/nature09710>.
57. Amodei, D, Hernandez, D. AI and Compute. OpenAI Blog; 2018. Retrieved Dec 10, 2020 from: <https://openai.com/blog/ai-and-compute/>
58. Graves A, Wayne G, Reynolds M, Harley T, Danihelka I, Grabska-Barwińska A, Colmenarejo SG, Grefenstette E, Ramalho T, Agapiou J, Badia AP, Hermann KM, Zwols Y, Ostrovski G, Cain A, King H, Summerfield C, Blunsom P, Kavukcuoglu K, Hassabis D. Hybrid computing using a neural network with dynamic external memory. *Nature*. 2016;538(7626):471–6. <https://doi.org/10.1038/nature20101>.
59. Huyck, CR. The neural cognitive architecture. AAAI 2017 fall symposium: technical report FS-17-05; 2017.
60. Epstein, SL. Navigation, cognitive spatial models, and the mind. AAAI 2017 fall symposium: technical report FS-17-05; 2017.

61. Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ. Building machines that learn and think like people. *Behav Brain Sci.* 2017;40:e253. <https://doi.org/10.1017/S0140525X16001837>.
62. Hawkins J, Lewis M, Klukas M, Purdy S, Ahmad S. A framework for intelligence and cortical function based on grid cells in the neocortex. *Front Neural Circuits.* 2019;12:121. <https://doi.org/10.3389/fncir.2018.00121>.
63. Laird, J, Mohan, S. Learning fast and slow: levels of learning in general autonomous intelligent agents. The Thirty-Second AAAI Conference on Artificial Intelligence (*AAAI* 2018). April 2018. Accessed at (Dec 10 2020): <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17261/16424>
64. Taatgen, NA. Cognitive architectures: innate or learned? AAAI 2017 fall symposium: technical report FS-17-05; 2017.
65. Goyal, A, Bengio, Y. Inductive biases for deep learning of higher-level cognition. arXiv preprint arXiv:2011.15091; Dec 7, 2020.



Mathematical Foundations of AIM

3

Yonina C. Eldar, Yuelong Li, and Jong Chul Ye

Contents

Introduction	38
Classical Machine Learning and Its Limitations	39
The Perceptron Model	39
Support Vector Machine	39
Modern Deep Learning Revolution	41
Architectures of Modern Deep Neural Networks	41
Representation Power of Deep Neural Networks	44
Other Properties of Deep Neural Networks	45
Algorithm Unrolling: From Iterative Algorithms to Deep Networks	46
Learned Iterative Shrinkage and Thresholding Algorithm	46
Unrolling Generic Iterative Algorithms	48
Interpretations of Deep Learning	49
Hierarchical Features in the Visual System	50
Geometric Understanding of Deep Neural Networks	51
Summary and Outlook	52
References	52

Abstract

With the tremendous success of deep learning in recent years, the field of medical imaging has undergone unprecedented changes. Despite the great success of deep learning in medical imaging, these recent developments are largely empirical. Our goal in this chapter is to provide an overview of some of the key mathematical foundations of deep learning to the medical imaging community. In particular, we will consider ties with traditional machine learning methods, unrolling techniques which connect deep learning to iterative algorithms, and

Y. C. Eldar (✉)
Department of Math and Computer Science, Weizmann Institute of Science, Rehovot, Israel
e-mail: yonina.eldar@weizmann.ac.il

Y. Li
Amazon, San Jose, CA, USA

J. C. Ye
Department Bio and Brain Engineering & Department Mathematical Sciences, Korea Advanced Institute of Science & Technology (KAIST), Daejeon, Republic of Korea
e-mail: jong.ye@kaist.ac.kr

geometric interpretations of modern deep networks.

Keywords

Machine learning · Deep learning · Neural network · Representation power · Hierarchical feature extraction

Introduction

Since groundbreaking performance improvements were first demonstrated by AlexNet [1] at the ImageNet challenge, deep learning has provided significant gains over classical approaches in various fields of artificial intelligence and data science. Availability of large-scale training datasets and advances in neural network research such as development of effective network architectures and efficient training algorithms have resulted in unprecedented successes of deep learning in innumerable medical imaging applications such as disease diagnosis, image segmentation, and image reconstruction.

In contrast to classical shallow machine learning approaches, which require “feature engineering” to extract features to feed into simple classifiers for diagnosis and recovery, one of the most important advantages of deep learning is that deep neural networks automatically discover the features and design appropriate classifiers and recovery methods in a data-driven way. This greatly simplifies the workflow of machine learning algorithm development and deployment. More importantly, the features are optimized toward specific tasks, which generally offer enhanced representation of the underlying data. In particular, for classification tasks the learned features can be much more discriminative than handcrafted features, whereas for reconstruction problems the learned features may better preserve the details and enable a more faithful reconstruction. As data representation plays a critical role, better features typically lead to superior performance in practice. Therefore, deep learning has become an increasingly important and versatile tool for medical imaging.

Nonetheless, the success of deep learning largely remains a mystery. From an architecture

perspective, deep neural networks are typically composed of a series of convolution, pooling, and nonlinearity layers, which from a mathematical point of view are regarded as primitive tools. Interestingly, with abundant training samples available, the cascaded connection of these primitive tools results in superior performance over traditional approaches which are carefully designed and tailored toward specific applications.

A popular explanation for the success of deep neural networks is that neural networks are developed by mimicking the human brain. One of the most famous numerical experiments is the emergence of hierarchical features from a deep neural network when it is trained to classify human faces [2]. This phenomenon is similarly observed in human brains, where hierarchical features of the objects emerge during visual information processing. However, when asked why, it is surprising to find out that neuroscientists usually rely on numerical simulations with artificial neural networks to explain how hierarchical properties arise in the brain [3].

To understand this fundamental question, one can go back to classical approaches to understand the similarities and differences from modern deep neural network methods. Recent studies have shown that there is a close relationship between deep learning approaches and sparse representations [4–8]. Specifically, neural networks have been interpreted as unrolled versions of sparse recovery, where each unfolded block is learned from the training data [4, 9–11]. The authors in [5, 6] showed that a deep neural network can be interpreted as a piecewise linear representation, whose data-driven basis is learned from training data and automatically adapted to various input signals [12]. In this chapter, we will consider these and other mathematical underpinnings of deep networks in order to offer insights into their unprecedented performance.

We begin in section “[Classical Machine Learning and Its Limitations](#)” by reviewing typical machine learning models and explain their limitations in representation power which motivate the development of modern deep learning techniques. We then review standard modern deep neural networks and illustrate conceptually how they have successfully achieved superior representation power compared to classical approaches in section “[Modern Deep Learning Revolution](#).[“Modern Deep Learning Revolution.”](#)

We next

review algorithm unrolling in section “[Algorithm Unrolling: From Iterative Algorithms to Deep Networks](#)” which connects traditional iterative algorithms with modern deep neural networks. To gain further insights into deep learning, we discuss how it can be interpreted, both from a biological and a geometric perspective in section “[Interpretations of Deep Learning](#). ” Finally, we summarize this chapter in section “[Summary and Outlook](#). ”

Classical Machine Learning and Its Limitations

In this section, we provide historical context on how deep learning evolved into its current form. We review two relevant classical machine learning models, the perceptron and support vector machine, and illustrate the limitations of these “shallow” models.

The Perceptron Model

One of the earliest machine learning models is the single layer perceptron [13]. As illustrated in Fig. 1, it is built by fully connected neurons at a single hidden layer, where each neuron is formed by an affine transformation of the input vector followed by a nonlinear mapping. Formally speaking, let $\varphi : \mathbb{R} \mapsto \mathbb{R}$ be a nonlinear activation function, and let $\chi \subset \mathbb{R}^n$ denote the input space. Then, a single layer perceptron $f_\Theta : \chi \mapsto \mathbb{R}$ is represented by

$$f_\Theta(x) = \sum_{i=1}^d v_i \varphi(w_i^\top x + b_i), \quad x \in \chi \quad (1)$$

where $w_i \in \mathbb{R}^n$ is a weight vector, $v_i, b_i \in \mathbb{R}$ are real constants, and $\Theta = \{(w_i, v_i, b_i)\}_{i=1}^d$ collectively represents the model parameters. To estimate the unknown parameters, a collection of training data $\{(x_i, y_i)\}_{i=1}^N$. The model parameters are then estimated by solving the following error criterion:

$$\min_{\Theta} \sum_{i=1}^N \ell(y_i, f_\Theta(x_i)) + \lambda R(\Theta) \quad (2)$$

where $\ell(.,.)$ is a desired loss function, λ is a regularization parameter, and $R(\Theta)$ is a regularization function with respect to the parameter set Θ .

When the single layer perceptron was first introduced, it was not clear how to simultaneously optimize the weights $\{w_i\}_{i=1}^d$ for all neurons. Instead, a heuristic algorithm called *the perceptron algorithm* was used to estimate the neuron weights. When the training data set is linearly separable, the perceptron algorithm is guaranteed to find an exact separation in a finite number of steps. Later on, the back-propagation algorithm was introduced [14], which applies gradient-based learning to learn all neuron weights simultaneously.

One of the classical results regarding the representation power of a single layer perceptron is the universal approximation theorem [15], which states that a feed-forward network with a single hidden layer containing a finite number of neurons can approximate continuous functions on compact subsets under mild assumptions on the activation function. The universal approximation theorem promoted research into neural networks as a powerful functional approximation; however, it turned out to be a limitation in the development of machine learning by circumventing construction of deeper neural networks. Although the theorem conceptually justifies that a shallow network with sufficiently many neurons can be a universal approximator, the proof of the theorem does not offer even a loose upper bound on how many neurons are required. Therefore, it is entirely possible that the networks which act as universal approximators are too large to be practical. However, at the time, deeper neural networks were difficult to train. Therefore, the prevalent approach was to increase the width of the network, namely the number of nodes, rather than the depth. Only recently has it been realized that depth matters, i.e., there exists a function that a deep neural network can approximate but a shallow neural network with the same number of parameters cannot [16–19].

Support Vector Machine

Another typical example of a “wide” model is Support Vector Machine (SVM). Similar to the neuron model, it classifies the data samples by learning an optimal separating hyperplane. However, in order to ensure that all data samples are far apart from the hyperplane, it employs a maximal

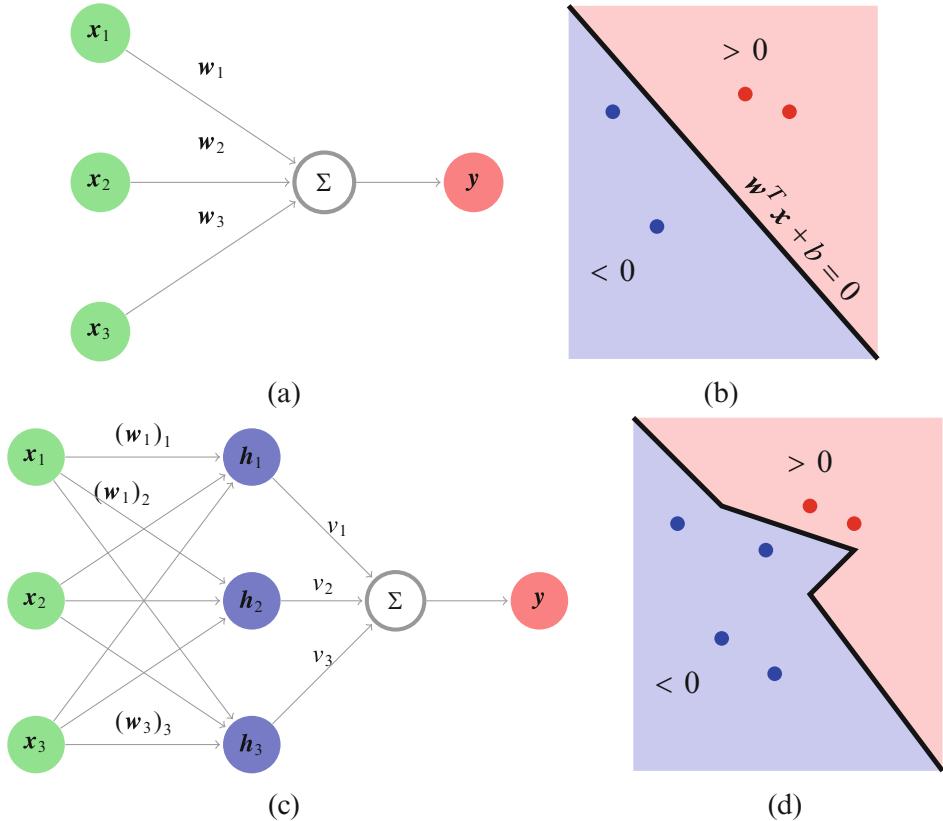


Fig. 1 Geometric interpretation of the perceptron model. (a) Architecture of the perceptron model for a three dimensional input. A neuron first applies an affine mapping of the input x and then performs thresholding to determine the predicted class y ; (b) geometrically, a neuron divides the

input space into two decision regions through a separating hyperplane; (c) the perceptron model is formed through a combination of multiple neurons; and hence (d) its decision boundary corresponds to a piecewise linear surface

margin loss function for training the weights of the separating hyperplane. Specifically, given training samples $\{x_n, y_n\}_{n=1}^N$, it solves the following minimax problem [20]:

$$\max_{w, b} \left\{ \frac{1}{\|w\|} \min_n [y_n(w^T x_n + b)] \right\} \quad (3)$$

which can be interpreted as maximizing the distances (margin) from the closest data samples (support vectors) to the separating hyperplane. A visual illustration of this criterion is given in Fig. 2a.

In practice, it is possible that the data samples cannot be classified by a single separating

hyperplane, i.e., the data samples are not linearly separable. In this scenario, SVM employs a non-linear mapping ϕ which embeds the data manifold into a high-dimensional feature space [21], where the data samples are assumed linearly separable. In practice, instead of explicitly designing the embedding ϕ , one can seek a *kernel function*, which characterizes the inner product $\langle ., . \rangle$ in the underlying feature space:

$$k(x, x') = \langle \phi(x), \phi(x') \rangle. \quad (4)$$

The aforementioned SVM technique is then employed in the feature space. In this way, the explicit form of ϕ does not need to be specified and only the kernel function has to be designed.

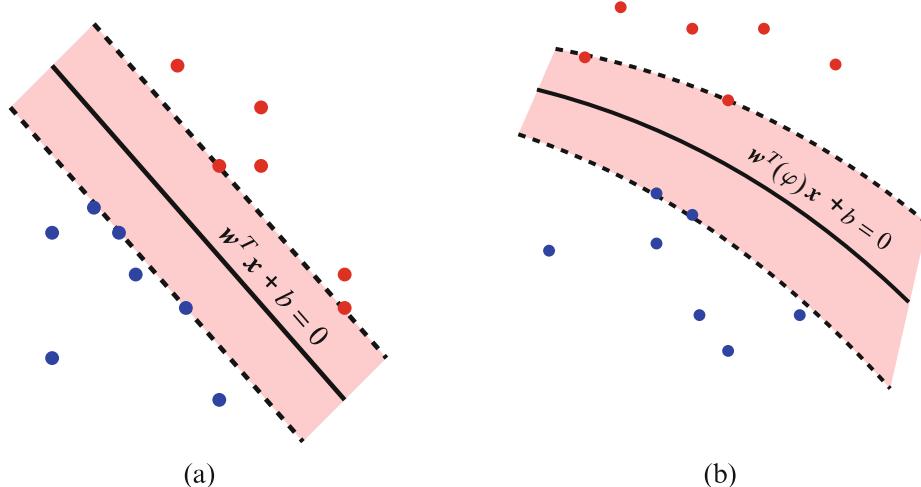


Fig. 2 The decision regions of SVM. (a) Linear SVM forms its separating hyperplane by maximizing the margins of data samples; (b) through the kernel trick, SVM is

This technique is commonly called the *kernel trick* [21]. In order for the kernel function to correspond to a valid inner product, Mercer’s theorem [22] requires φ that the kernel function is symmetric and positive definite. As depicted in Fig. 2b, the kernel trick enables SVM to form nonlinear decision boundaries. As the feature space is often of higher dimensionality compared to the ambient space, kernel SVM can be regarded as increasing the “width” of the classification model in order to enhance its capacity, instead of employing hierarchical architectures which follow the “depth” dimension.

In principle, any symmetric positive definite kernel function can be associated to an inner product. Therefore, the kernel generates a wide variation of functions within the feature space. Indeed, one of the main research thrusts in classical machine learning approaches is to find appropriate kernels that are suitable for specific applications. That said, kernel methods still have fundamental limitations. First, the kernel function is typically handcrafted instead of learned from data. Second, once the kernel machine is trained, the parameters are fixed, and it is not possible to adjust them at test phase. These drawbacks lead to fundamental limitations of expressivity of kernel-based learning models [21].

able to form nonlinear decision boundaries and classify samples that are not linearly separable

Modern Deep Learning Revolution

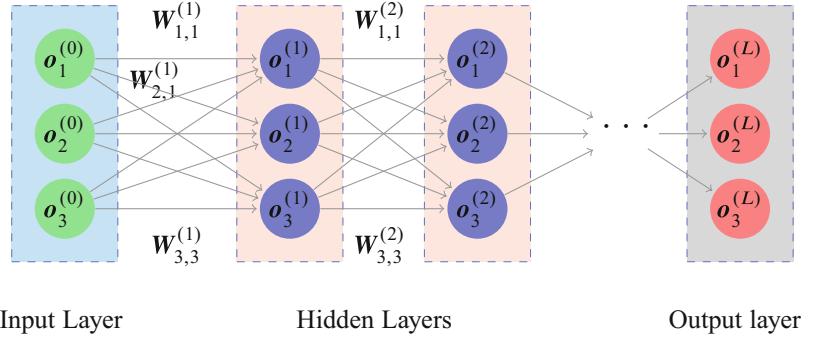
One of the main reasons for the success of deep learning is its significantly extended expressivity by learning hierarchical features through deep neural networks. Before we delve deep into this claim, we first provide a brief review of modern deep network architectures.

Architectures of Modern Deep Neural Networks

In early neural network research, the Multi-Layer Perceptrons (MLP) was a popular choice. A diagram illustration of this architecture is given in Fig. 3. As can be observed, this model can be regarded as the extension of a single layer perceptron shown in Fig. 1c, by introducing multiple hidden layers. This scheme can be viewed as adopting a hierarchical feature representation and increasing its “depth.”

Similar to perceptron, each neuron in MLP is fully connected to every neuron in the previous layer, except for the input layer. The layers are thus commonly called *fully connected* layers. Analytically, in the l -th layer, the

Fig. 3 Architecture of MLP. The input vector passes through a few hidden layers and reaches the output layer. Each layer comprises an affine transformation followed by a nonlinear activation function. We omit drawing the activation functions for brevity



relationship between the neurons $o_j^{(l)}$ and $o_i^{(l+1)}$ is expressed as

$$o_i^{(l+1)} = \sigma \left(\sum_j w_{ij}^{(l+1)} o_j^{(l)} + b_i^{(l+1)} \right) \quad (5)$$

where $W^{(l+1)}$ and $b^{(l+1)}$ are the *weights* and *biases*, respectively, and σ is a nonlinear *activation function*. Popular choices of activation functions include the logistic function and the hyperbolic tangent function. In recent years, they have been superseded by Rectified Linear Units (ReLU) [23] defined by

$$\text{ReLU}(x) = \max \{x, 0\}. \quad (6)$$

The W's and b's are generally trainable parameters that are learned from datasets, using back-propagation [24] for gradient computation.

Nowadays, MLPs are rarely seen in practical medical applications. The fully connected nature of MLPs contributes to a rapid increase in their parameters, making them difficult to train. To address this limitation, Fukushima et al. [25] designed a neural network by mimicking the visual nervous system [26]. The neuron connections are restricted to local neighbors only, and weights are shared across different spatial locations. The affine transformations then become convolutions (or correlations in a strict sense), and the resulting networks are commonly called Convolutional Neural Networks (CNN). A visual illustration of a CNN can be seen in Fig. 4. With significantly reduced parameter dimensionality, training deeper networks becomes much easier.

While CNNs were first applied to digit recognition, their translation invariance is a desirable

property for analyzing image features and are broadly applied in various medical problems, including medical image retrieval [27], segmentation [28], and reconstruction [4], to name a few. In reality, the architecture of CNN can be more sophisticated than illustrated in Fig. 4. There may be advanced types of convolutions, such as stridden convolutions which reduce spatial resolution [29], transposed convolutions which perform up-sampling [30], and more. There may also be other layers, such as pooling layers which perform spatial aggregation [29], batch normalization layers which stabilize training [31], and dropout layers which perform network assembling to reduce overfitting [32].

In some medical applications, the data may exhibit certain sequential forms. A concrete example is video processing, where video frames have temporal dependencies [33]. In such scenarios, Recurrent Neural Networks (RNN) [14] are a popular choice. RNNs explicitly model the sequential data dependence in different time steps in the sequence and scale well to sequences with varying lengths. A visual depiction of RNNs is provided in Fig. 5. Given the previous hidden state $s^{(l-1)}$ and input variable $x^{(l)}$, the next hidden state $s^{(l)}$ is computed as

$$s^{(l)} = \sigma_1 \left(W s^{(l-1)} + U x^{(l)} + b \right) \quad (7)$$

while the output variable $o^{(l)}$ is generated by

$$o^{(l)} = \sigma_2 \left(V s^{(l)} + b \right)$$

Here, U , V , W , and b are trainable network parameters and σ_1 and σ_2 are activation

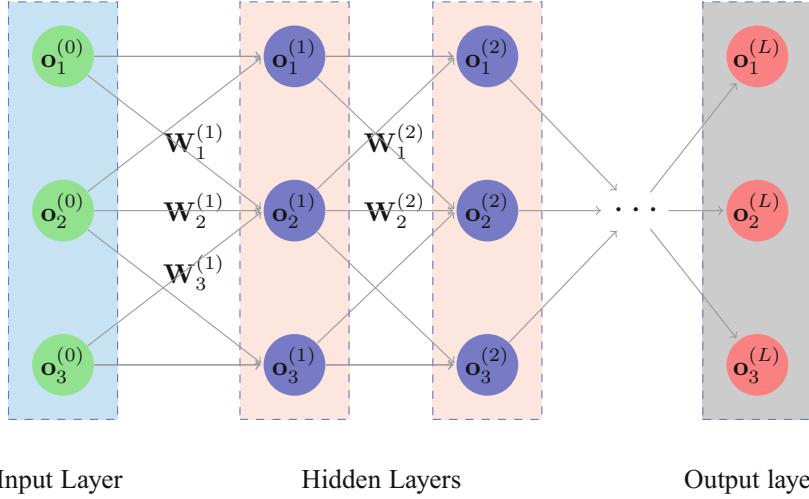


Fig. 4 Architecture of CNN. Instead of connecting all neurons in adjacent layers, CNN only connects each neuron to its spatial neighbors in the previous layer. Furthermore, the weights are shared across different spatial

locations. The output neurons in each layer can then be generated by convolving the input neurons with weights, followed by a nonlinear activation function. We omit drawing the activation functions for brevity

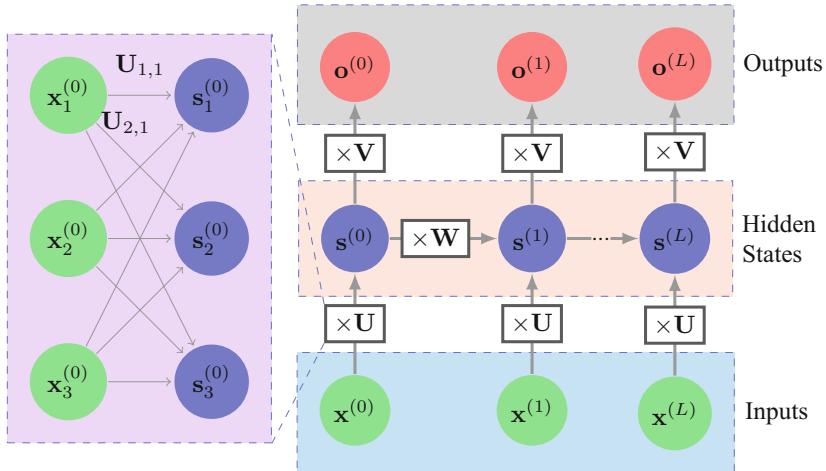


Fig. 5 Architecture of RNN. The inputs $\{x^{(t)}\}_1$ are fed into the network in a sequential order. The state variables $\{s^{(t)}\}_1$ evolve according to a linear transition model, thus capturing sequential dependencies. The outputs $\{o^{(t)}\}_1$ are

generated by combining the state variables and the inputs. Parameters U , V , and W are shared across different time steps

functions. We again omit the activation functions and biases in Fig. 5. In contrast to MLPs and CNNs where the layer operations are applied recursively in a hierarchical representation fashion, RNNs apply the recursive operations as the time step evolves. A distinctive property of RNNs is that the parameters U , V , and W are shared across all time steps, rather than varying from layer to layer. Training

RNNs can thus be difficult as the gradients of the parameters may either explode or vanish [34].

In practice, the state-space relation (7) suffers from a few limitations: It does not favor long-term dependencies which is crucial for modeling long sequences, and it brings about difficulties in training by introducing the gradient vanishing and exploding phenomena. To address these issues,

more advanced architectures have been suggested. Long-Short-Term-Memory [35] and Gated Recurrent Network [36] employ gating units and allow information to flow freely both in the forward pass and during back-propagation, which effectively mitigates the aforementioned limitations of vanilla RNN. Recently, self-attention mechanisms have become popular as an effective structure to reduce computational complexity and enable parallel computations, and better capture long-term dependencies [37]. However, the basic prototype remains the same: The input sequence is first encoded into state vectors, and then decoded into an output sequence. The decoder models the temporal dependencies between the current and previous timestamps, which can be regarded as an auto-regressive model.

Representation Power of Deep Neural Networks

In order to understand the superior representation power of modern deep neural networks over classical machine learning models, we will explore the decision regions of MLP by analyzing its per-layer mapping (5). Without loss of generality, we omit the bias term. For the nonlinear activation function σ , we adopt the ReLU function defined in (6), and let

$$o^{(l)} = \sigma(g^{(l)}) , \quad g^{(l)} := W^{(l)} o^{(l-1)}. \quad (8)$$

For an L -layer MLP, the neural network output for a given input x can be represented by

$$f_{\Theta}(x) := (\sigma \circ g^{(L)} \circ \sigma \circ g^{(L-1)} \dots \circ g^{(1)})(x) = B(x)^T x \quad (9)$$

where $0 = [W^{(1)} \dots W^{(L)}]$ and

$$B(x) = W^{(1)} \Lambda^{(1)}(x) W^{(2)} \Lambda^{(2)}(x) \dots \Lambda^{(L-1)}(x) W^{(L)} \quad (10)$$

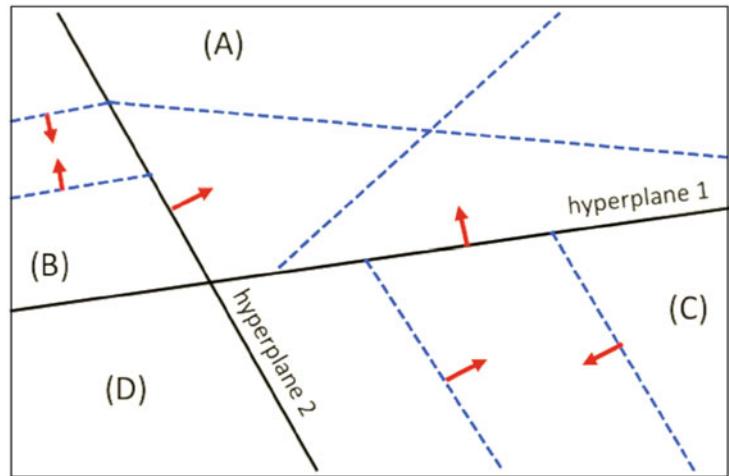
Here, $\Lambda^{(l)}(x)$ is a diagonal matrix with 0 and 1 elements indicating the ReLU activation patterns.

Note that the matrix $B(x)$ in (10) depends on the ReLU activation patterns $\Lambda^{(l)}, l = 1, \dots, L - 1$, which are determined by the input x . In fact, this ReLU activation-dependent diagonal matrix plays a key role in enabling inductive learning and emergence of hierarchical features. Specifically, the nonlinearity is applied after the linear operation, so the on-and-off activation pattern of each ReLU determines a binary partition of the feature space at each layer across the hyperplane that is determined by the weight matrix. Accordingly, in deep neural networks, the input space is partitioned into multiple nonoverlapping regions so that input images for each region share the same linear representation, but not across the partition. This implies that two different input images are automatically switched to two distinct linear representations that are different from each other depending on the partition as shown in Fig. 6 [38].

This leads to an important insight: Deep neural networks search for the distinct linear representation for each input. However, in contrast to classical optimization-based approaches, deep neural networks do not solve the optimization problem for a new input, rather they switch to different linear representations by changing the ReLU activation patterns. This is an important advance over the classical signal-processing approach.

To quantify the representation power of modern deep networks such as CNN, researchers are conducting rigorous theoretical analysis to extend the classical universal approximation theorem. Zhou et al. [39] prove that deep CNNs can approximate continuous functions supported on a compact space of any accuracy as long as its depth is sufficiently high. They also analyze the approximation property of deep CNNs over a family of functions in Sobolev space, which, loosely speaking, comprises smooth functions that are differentiable to a certain order. Similarly, Yarotsky et al. [19] analyze the approximation property of deep neural networks with ReLU activation functions on the Sobolev space. In addition, they provide upper and lower bounds of network complexity, i.e., the number of networks layers, weights, and neurons required to reach a given approximation error.

Fig. 6 Piecewise linear representation by a deep neural network



Other Properties of Deep Neural Networks

Splines, or piecewise polynomials, are another form of function approximators. It is therefore interesting to see how splines are connected to deep neural networks. Unser et al. [12] establish a conceptual connection between deep neural networks and piecewise linear functions, i.e., first-order uniform splines. Suppose we are given an MLP of the following form:

$$f\left(x; \left\{W^{(l)}, \sigma^{(l)}\right\}_l\right) := \left(\sigma^{(L)} \circ g^{(L)} \circ \sigma^{(L-1)} \circ g^{(L-1)} \dots \circ g^{(1)}\right)(x)$$

where $g^{(l)} := W^{(l)} o^{(l-1)}$ and $\sigma^{(l)}$'s are arbitrary nonlinear activation functions.

Given a collection of training samples $\{\mathbf{x}_m, \mathbf{y}_m\}_m$, we train the network f through the following optimization problem:

$$\begin{aligned} \min_{\{W^{(l)}, \sigma^{(l)}\}} & \sum_{m=1}^M \ell\left(y_m, f\left(x_m; \left\{W^{(l)}, \sigma^{(l)}\right\}_l\right)\right) \\ & + \lambda \mathcal{R}\left(\left\{W^{(l)}\right\}_l\right) + \mu \mathcal{T}\left(\left\{\sigma^{(l)}\right\}_l\right) s \mathcal{T} \end{aligned} \quad (11)$$

where ℓ is a general convex loss function, \mathcal{R} is some arbitrary convex regularization function over the network weights, \mathcal{T} is a regularization over $\sigma^{(l)}$'s which promotes sparsity of their derivatives, and λ and μ are positive regularization parameters. The authors prove that, in order for f to be an optimal solution to the problem (11), it

must be the case that $\sigma^{(l)}$'s are piecewise linear. In other words, the original infinite-dimensional optimization becomes finite-dimensional because $\sigma^{(l)}$ now adopts a parametric spline form.

Another intriguing property of deep neural networks is their amazing generalization capability, which seems mysterious from the perspective of classic machine learning. In particular, the number of trainable parameters in deep neural networks is often greater than the training data set, this situation being notorious for overfitting from the point of view of classical statistical learning theory. However, empirical results have shown that a deep neural network generalizes well in the testing phase, resulting in high performance for the unseen data.

This apparent contradiction has raised questions about the mathematical foundations of machine learning and their relevance to practitioners. Recently, the authors in [40, 41] have suggested how to reconcile classical understanding and modern practice in a unified framework. In classical machine learning theory, models with exceedingly high capacity are subject to overfitting and exhibit high test errors due to the fundamental bias-variance trade-off. However, the authors argue that, once the model capacity increases beyond a certain point called interpolation point, its performance starts improving in the test phase. Increasing the functional class capacity to the overparameterized area thus improves the generalization performance of the resulting

classifiers. The authors further justify their claim empirically through experiments.

Algorithm Unrolling: From Iterative Algorithms to Deep Networks

Parallel to the revolution of machine learning models, another thread of research constructs deep networks from iterative algorithms by mapping each iteration to a network layer. This line of research centers around an important technique called *algorithm unrolling*, which originated from Gregor et al.'s seminal technique on learnable sparse coding, called Learned Iterative Shrinkage and Thresholding Algorithm (LISTA) [7]. We first review this method and related theoretical findings. We then discuss general formulations of algorithm unrolling and provide practical examples.

Learned Iterative Shrinkage and Thresholding Algorithm

LISTA aims to approximately solve the *sparse coding* problem at a higher efficiency than iterative methods, by learning an unknown dictionary from real data. Specifically, given an observation vector $y \in \mathbb{R}^m$, we seek a vector $x \in \mathbb{R}^n$ such that $y \approx Wx$ and encourage as many coefficients in x to be zero (or small in magnitude) as possible [42]. A typical approach to achieve this is by solving an unconstrained convex minimization problem:

$$\min_{x \in \mathbb{R}^m} \frac{1}{2} \|y - Wx\|_2^2 + \lambda \|x\|_1, \quad (12)$$

where $\lambda > 0$ is a regularization parameter that controls the sparsity of the solution. A well-known class of methods for solving (12) are proximal methods such as Iterative Shrinkage and Thresholding Algorithm (ISTA) [43], which perform the following iterations:

$$x^{l=1=s_\lambda} \left\{ \left(I - \frac{1}{\mu} W^T W \right) x^1 + \frac{1}{\mu} W^T y \right\}, l = 0, 1, \dots \quad (13)$$

Here, $I \in \mathbb{R}^{n \times n}$ is the identity matrix, μ is a positive parameter that controls the iteration step size, $S\lambda(\cdot)$ is the soft-thresholding operator defined as for a scalar x , and $S\lambda(-)$ operates element-wise on vectors and matrices.

$$S\lambda(x) = \text{sign}(x) \max \{|x| - \lambda, 0\}, \quad (14)$$

The slow convergence rate of ISTA can be problematic in real-time applications. Furthermore, the matrix W may not be known exactly. In their seminal work, Gregor and Lecun [7] propose a highly efficient learning-based method that computes good approximations of optimal sparse codes in a fixed amount of time, with the help of W learned optimally from real data [7]. Specifically, iteration (13) can be recast into a single network layer as depicted in Fig. 7. This layer comprises a series of analytical operations (matrix-vector multiplication, summation, and soft-thresholding), which is of the same nature as a neural network. A diagram representation of one iteration step reveals its resemblance to a single network layer. Executing ISTA L times can be interpreted as cascading L such layers, which essentially forms an L -layer deep network. Note that, in the unrolled network, an implicit substitution of parameters has been made: $W_t = I - \frac{1}{\mu} W^T W$ and $W_e = \frac{1}{\mu} W^T$.

After unrolling ISTA into a network, named Learned ISTA (LISTA), the network is trained through back-propagation using real datasets to optimize the parameters W_b , W_e , and λ . Training is performed in a supervised manner, meaning that for every input vector $y^t \in \mathbb{R}$, $t = 1, \dots, T$, its corresponding sparse output $x^{*t} \in \mathbb{R}^n$, $t = 1, \dots, T$ is known. The sparse codes x^{*t} can be determined, for example, by executing ISTA when W is known. Feeding vector y^t into the network results in a predicted output $\tilde{x}^t(y^t; W_t, W_e, \lambda)$. The network-training loss function is formed by comparing the prediction with the known sparse output x^{*t} :

$$\ell(W_t, W_e, \lambda) = \frac{1}{T} \sum_{t=1}^T \| \tilde{x}^t(y^t; W_t, W_e, \lambda) - x^{*t} \|_2^2 \quad (15)$$

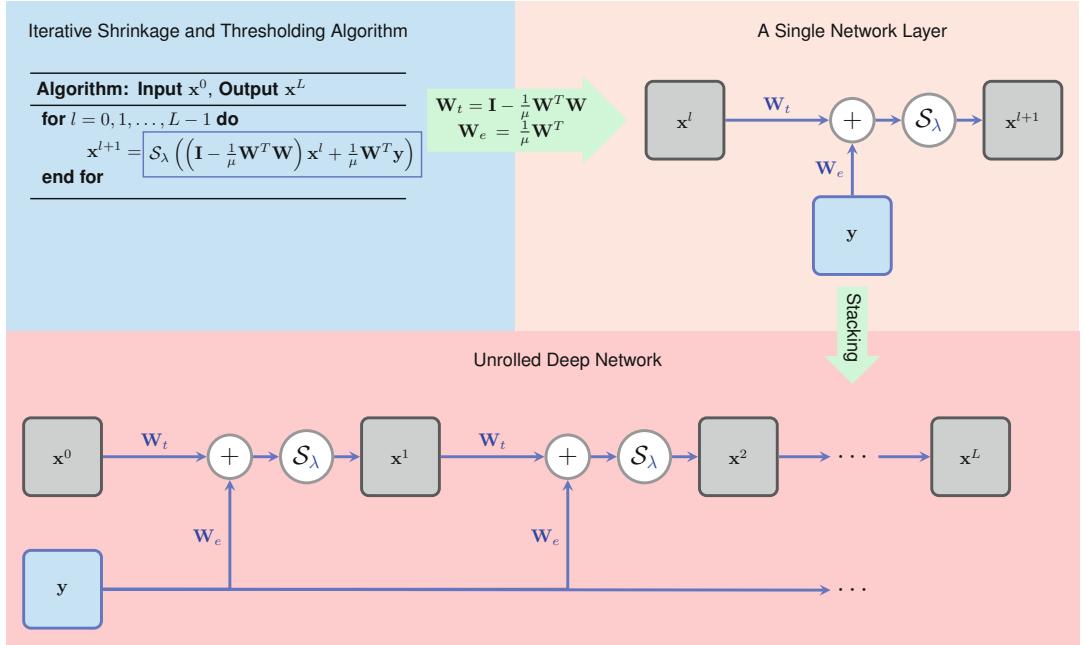


Fig. 7 Illustration of LISTA: One iteration of ISTA executes a linear and then a nonlinear operation and thus can be recast into a network layer; by stacking the layers, a deep network is subsequently

The network is trained through loss minimization, using popular gradient-based learning techniques such as stochastic gradient descent, to learn the unknown parameters W_b , W_e , and λ [24]. With optimized parameters, LISTA may achieve higher efficiency compared to ISTA. Indeed, it has been shown empirically that the number of layers L in (trained) LISTA can be an order of magnitude smaller than the number of iterations required for ISTA to achieve convergence corresponding to a new observed input, thus dramatically boosting the sparse coding efficiency [7]. Furthermore, when the dictionary W_t and W_e are unknown or hard to determine analytically, they can be learned from real datasets. In practice, W_t , W_e , and λ can be untied and vary in each layer.

In addition to empirically observed superior efficiency, researchers have recently confirmed the faster convergence speed of LISTA over ISTA by conducting rigorous theoretical analysis. In particular, recent studies reveal the convergence rate of LISTA, and relevant techniques,

trained using paired inputs and outputs by back-propagation to optimize the parameters W_e , W_t , and λ . The trainable parameters in the network are colored in blue

and characterize the optimality conditions. For instance, Xin et al. [44] study the unrolled Iterative Hard Thresholding (IHT) algorithm, which has been widely applied in various sparsity-constrained estimation problems. IHT largely resembles ISTA except that an ℓ^0 norm is employed instead of the ℓ^1 norm. The authors prove that, in order for the unrolled IHT network to recover a maximally sparse solution (i.e., a vector with minimal ℓ^0 norm), a weight coupling scheme must be satisfied; furthermore, under the weight-coupling constraint and certain additional Restricted Isometry Property (RIP) conditions [11], a linear convergence rate can be deduced. Compared to classical IHT, the learned version poses a much milder requirement on the RIP condition, meaning that the unrolled network is capable of recovering sparse signals from a much broader family of dictionaries.

Chen et al. [45] observe similar behaviors of the LISTA network with layer-specific parameters. They prove that, in order for LISTA to

recover the underlying sparse solutions, a similar weight-coupling constraint must be satisfied asymptotically. They further introduce a so-called support-selection scheme, which, together with weight coupling and a few other mild conditions, ensures linear convergence rate of the unrolled network. As a follow-up, Liu et al. [46] introduce certain mutual coherence conditions and analytically characterize optimal network parameters based on those conditions. Similar to the networks with trained weights, networks adopting analytic weights converge at a linear rate, which implies that analytic weights can be as efficient as learned weights. In addition, analytic weights are of much lower dimensionality compared to trained weights. However, determining the analytic weights can be a nontrivial task as they are solutions to another dedicated optimization problem.

Unrolling Generic Iterative Algorithms

Although the initial focus of Gregor et al.'s work [7] was on sparse coding techniques, the underlying principles could be easily generalized. More specifically, provided with a certain iterative algorithm, we can unroll it into a corresponding deep network, following the procedures depicted in Fig. 8 [8]. The first step is to identify the analytic

operations per iteration, which we represent abstractly as an h function, and the associated parameters, which we denote collectively as θ^l . The next task is to generalize the functional form of h into a more generic version \hat{h} , and correspondingly expand the parameters θ^l into an enlarged version $\hat{\theta}^l$ if necessary. For instance, in LISTA, the parameter W is substituted with W_t and W_e through the formula $W_t = I - \frac{1}{\mu} W^T W$ and $W_e = \frac{1}{\mu} W^T$. After this procedure, each iteration can be recast into a network layer in the same spirit as LISTA. By stacking the mapped layers together, we obtain a deep network with undetermined parameters and then obtain optimal parameters through end-to-end training using real-world datasets.

The exact approach to generalize h and θ^l 's toward \hat{h} and $\hat{\theta}^l$'s is largely case specific. An extreme scenario is to strictly follow the original functional forms and parameters, i.e., to take $h = \hat{h}$ and $\theta^l = \hat{\theta}^l, \forall l$. In this way, the trained network corresponds exactly to the original algorithm with finite truncation and optimal parameters. In addition to efficiency enhancement thanks to training [7], the unrolled networks can aid with estimating structured parameters such as filters [47] or dictionaries [48] which are hard to design either analytically or by handcrafting. Alternatively, some operations may be replaced with a stand-

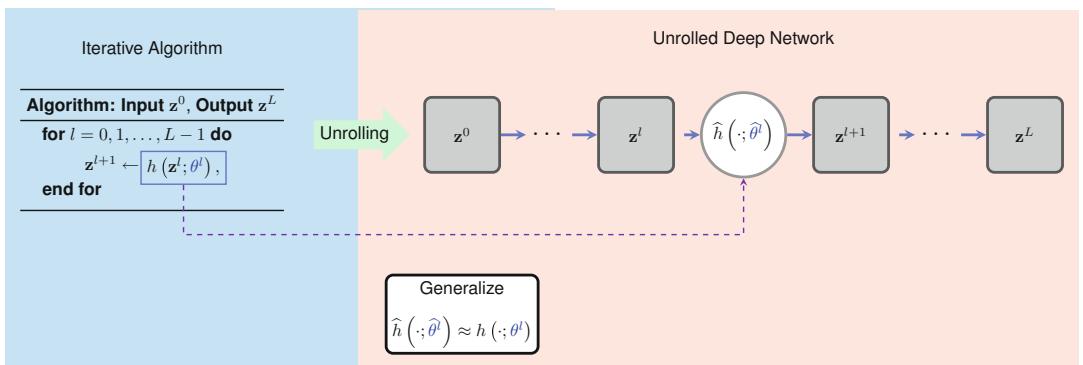


Fig. 8 Illustration of the general idea of algorithm unrolling: given an iterative algorithm, we map one iteration (described as the function h parametrized by $\theta^l, l = 0, \dots, L-1$) into a single network layer, and stack a finite number of layers to form a deep network. Feeding the data forward through an L -layer network is equivalent to

executing the iteration L times (finite truncation). The parameters $\theta^l, l = 0, \dots, L-1$ are learned from real datasets by training the network end-to-end to optimize the performance. They can either be shared across different layers or varying from layer to layer. The trainable parameters are colored in blue

alone deep neural network such as Convolutional Neural Network (CNN) or Recurrent Neural Network (RNN). For instance, in [49], the authors replace a proximal gradient update step with a CNN. In addition, the parameters can be layer specific instead of being shared across different layers. For instance, in [50], the authors plug in a CNN in each iteration step (layer) and allow the network parameters to differ. As it is, networks with shared parameters generally resemble RNN, while those with layer-specific parameters mimic CNN, especially when there are convolutional structures embedded in layer-wise operations. While custom modifications may potentially invalidate the convergence guarantees, they are practically beneficial and critical for performance improvement because the representation capacity of the network can be significantly extended.

In addition to performance and efficiency benefits, unrolled networks can potentially reduce the number of parameters and hence storage footprints. Conventional generic neural networks typically reuse essentially the same architectures across different domains and thus require a large amount of parameters to ensure their representation power. In contrast, unrolled networks generally carry significantly fewer parameters, as they implicitly transfer problem structures (domain knowledge) from iterative algorithms to unrolled networks, and their structures are more specifically tailored toward target applications. These benefits not only ensure higher efficiency, but also provide better generalizability especially under limited training schemes.

Unrolling techniques have been widely used in medical applications. An important imaging modality is ultrasound, which has the advantage of being a radiation-free approach. When used for blood flow depiction, one of the challenges is the fact that the tissue reflections tend to be much stronger than those of the blood, leading to strong clutter resulting from the tissue. Thus, an important task is to separate the tissue from the blood. Various filtering methods have been used in this context such as high-pass filtering, and filtering based on the singular value decomposition. Solomon et al. [51] suggest using a robust Principal Component Analysis (PCA) approach by

modeling the received ultrasound movie as a low-rank and sparse matrix where the tissue is low rank and the blood vessels are sparse. The robust PCA problem can be solved via a generalized version of ISTA, which is further unrolled into a deep network, called Convolutional rObust pRincipal cOmpoNent Analysis (CORONA). As the name suggests, they replace matrix multiplications with convolutional layers, effectively converting the network into a CNN-like architecture. Compared with state-of-the-art approaches, CORONA demonstrates vastly improved reconstruction quality and has much fewer parameters than the well-known ResNet [52].

In optical microscopy, a fundamental challenge is to enhance spatial resolution, which is limited by the physics of light. Solomon et al. [47] exploit the sparse nature of the fluorophores distribution and improve the spatial resolution via a sparsity-constrained estimation approach, called SPARSity based super-resolution COrrelation Microscopy (SPARCOM). Recently, Dardikman et al. [53] unroll SPARCOM into a deep network called Learned SPARCOM (LSPARCOM). The structure of the network resembles LISTA, except that they adopt a customized proximal operator. Experimental results show that LSPARCOM can obtain super-resolution images from a small number of high-emitter density frames without knowledge of the optical system, and has clear runtime advantages.

Interpretations of Deep Learning

We have so far covered the origin of deep learning and how it connects to classical machine learning models and traditional iterative algorithms. In this section, we investigate other ways to interpret deep neural networks, in addition to a methodological perspective. Specifically, we first take a biological perspective, where we illustrate how hierarchical features, an essential component of deep neural networks, are also commonly found in biological visual systems. We then switch to a geometric standpoint and contend that deep neural networks effectively capture the low-dimensional data manifolds.

Hierarchical Features in the Visual System

The visual system is a part of the central nervous system that enables organisms to detect and interpret information from visible light to create a representation of the environment. In pursuit of understanding the visual system, Hubel and Wiesel [54] found two classes of functional cells in the primary visual cortex: simple cells and complex cells. More specifically, simple cells at the primary visual cortex at the V1 L4 layer respond best to edge-like stimuli with a certain orientation, position, and phase within their relatively small receptive fields. They realized that such response of the simple cells could be obtained by pooling the activity of a small set of input cells with the same receptive field that is observed in Lateral Geniculate Nucleus (LGN) cells. This observation

has been extended to higher areas of the visual cortex to result in a class of object recognition models [3]. Specifically, there is a neuronal connection along this path, which forms a neuronal hierarchy such that neurons become sensitive to more complex inputs. An extreme form or surprising example of this information-processing hierarchy can be found in the discovery of the so-called “Jennifer Aniston Cell” [55], which identified a single neuron that is sensitive to a complex but specific concept or object.

A similar phenomenon can be observed in the convolution neural network, once it is properly trained. In particular, VGGNet [2] provides very intuitive information that is well correlated with the visual information processing in the brain. For example, Fig. 9 illustrates the input signal that maximizes the filter response at specific channels and layers of VGGNet [2]. Here, an input image

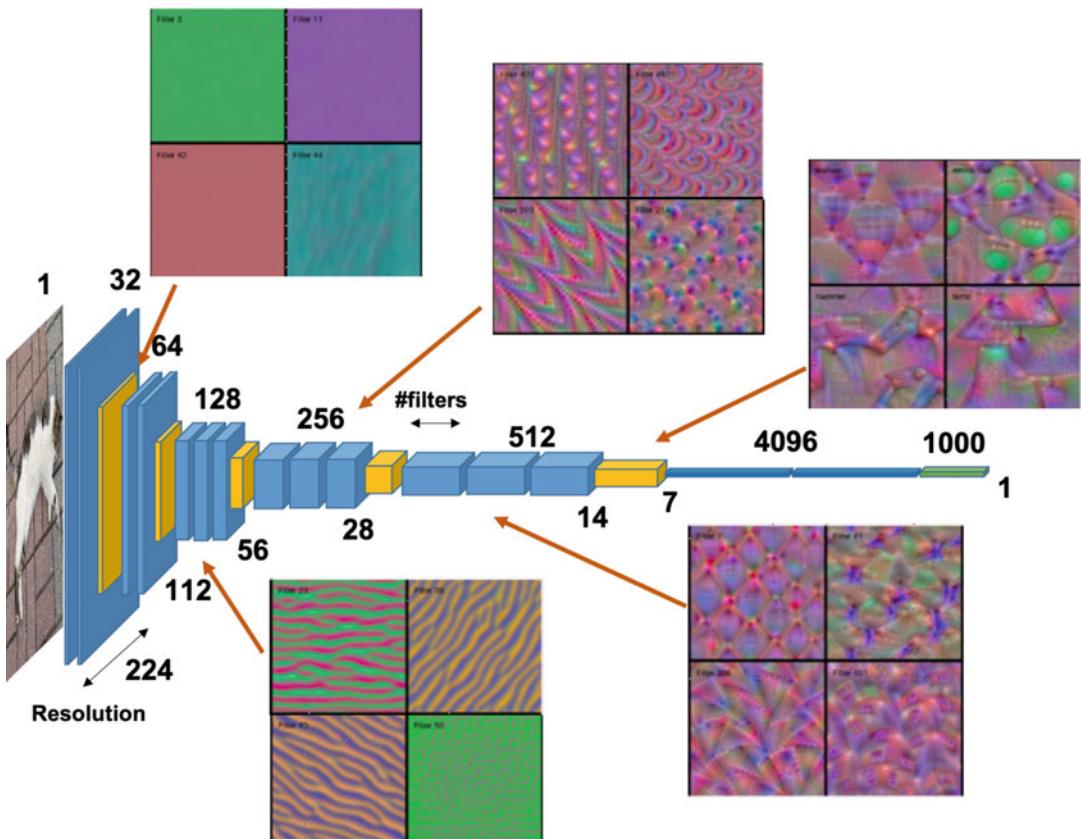


Fig. 9 Input images that maximize filter responses at specific channels and layers of VGGNet

that activates this filter most is displayed for specific channel and layer filters. This is similar to the Hubel and Wiesel experiments where they analyzed the input image that maximizes the neuronal activation. Specifically, at the earlier layers the input signal-maximizing filter response is composed of directional edges similar to the Hubel and Wiesel experiment. As we go deeper into the network, the filters build on each other and learn to code more complex patterns. Finally, in Fig. 9 the input images that maximize the response on the last softmax level in the specific classes correspond to the visualization of the input images that maximize the class categories. The emergence of the hierarchical feature from simple edges to the high-level concept is similar to visual information processing in the brain.

Geometric Understanding of Deep Neural Networks

The manifold structure of real-world data has been heavily exploited in classical machine learning techniques. Structured data are often assumed to lie on a manifold whose dimensionality is much lower than its ambient space. In particular, it has been long recognized that natural images lie on a low-dimensional manifold [34], and the key to

success in many machine learning tasks hinges on capturing the underlying manifold structure. As illustrated in Fig. 10, researchers have spent intensive efforts to model the bidirectional mapping between the data manifold and the underlying latent space. The forward mapping φ_α , commonly called the *encoder*, maps each data sample (such as an image) into a low-dimensional latent vector, while the inverse mapping $\varphi_{\alpha-1}$, commonly called *decoder*, generates a sample (such as an image) from a provided latent vector.

In classical machine learning, modeling the encoder and decoder has been a key topic in unsupervised learning. For instance, Principal Component Analysis (PCA), as a widely applied linear dimensionality reduction technique, estimates a linear encoder which projects the data onto a low-dimensional subspace, and a linear decoder which recovers the data approximately from this subspace. However, the subspace usually does not offer an accurate approximation when the data manifold is nonlinear. To improve the approximation accuracy, nonlinear dimensionality techniques have been developed by either modeling the topological relationship among inputs [56], or capturing their metric structure [57]. Nonetheless, these techniques lack a deep hierarchical decomposition of the mapping, φ_α , and hence do not faithfully model the data manifold.

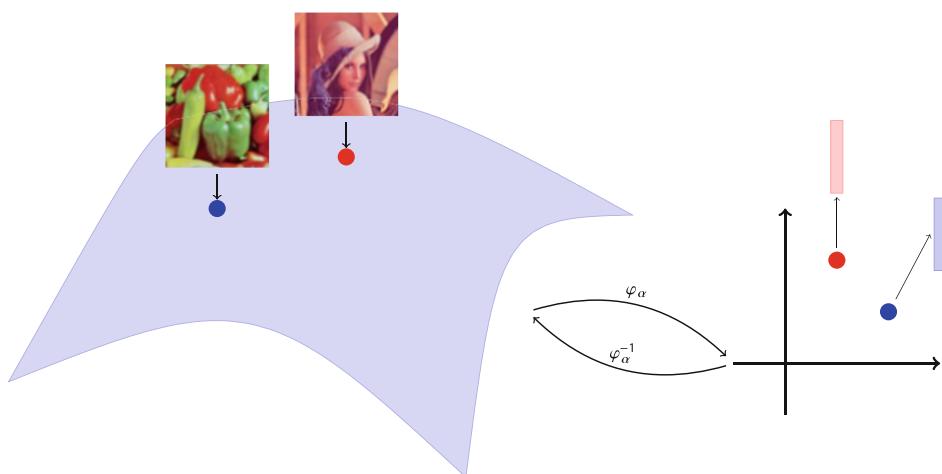


Fig. 10 Manifold structure of natural images. Many unsupervised learning techniques aim to discover the encoding mapping φ_α from the low-dimensional manifold

of natural images to its latent space. Generative models capture the inverse mapping, which generates natural image samples given the latent vectors

In recent years, deep unsupervised learning has achieved tremendous progress. The essence of many such techniques is how to learn the encoder/decoder using deep neural networks. Variational Auto Encoders (VAE) [58] jointly learns both the encoding mapping φ_α and the decoding mapping $\varphi_{\alpha-1}$ by integrating auto-encoder with variational Bayes. There are also generative models which effectively learn the mapping $\varphi_{\alpha-1}$ and generate high-quality data samples providing the latent vectors. Typical generative models include Generative Adversarial Networks (GAN) [59] and Normalizing Flows (NF) [60]. The deep hierarchical architecture ensures high capacity in modeling complicated functional mappings, which is the key to success for these approaches.

Summary and Outlook

In this chapter, we reviewed the historical developments of deep learning, following two main threads: the evolution from classical machine learning models to modern deep learning models, and the transition from traditional iterative algorithms to contemporary deep networks. We explained the limitations of traditional machine learning models and algorithms, in terms of expressivity, and discussed how deep learning successfully can overcome these limitations. We also reviewed recent theoretical breakthroughs which justify the superior representation power of deep networks and help in understanding their properties and behaviors.

Although there is already a rich body of research on the mathematical foundation of deep learning, we are still far from understanding the full mystery of deep learning. As it is, so far there is little knowledge on what weights are optimal in order for the networks to approximate a certain function. In addition, currently network training is largely empirical, and network performance largely relies on heuristics in training and hyperparameter tuning. In the context of medical imaging, developing more methods that are interpretable and robust is of key importance. With the many recent exciting advances and the many researchers entering this field, the next

decade will surely lead to further insights and methods in these directions.

References

- Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Adv Neural Inform Process Syst*. 2012;25:1097–1105.
- Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: International conference on learning representations, 2015.
- Riesenhuber M, Poggio T. Hierarchical models of object recognition in cortex. *Nat Neurosci*. 1999;2(11):1019–25.
- Jin KH, McCann MT, Froustey E, Unser M. Deep convolutional neural network for inverse problems in imaging. *IEEE Trans Image Process*. 2017;26(9):4509–22.
- Ye JC, Han Y, Cha E. Deep convolutional framelets: a general deep learning framework for inverse problems. *SIAM J Imag Sci*. 2018;11(2):991–1048.
- Ye JC, Sung WK. Understanding geometry of encoder-decoder CNNs. *Int Conf Mach Learn*, 2019;97:7064–7073.
- Gregor K, LeCun Y. Learning fast approximations of sparse coding. *Int Conf Mach Learn*, 2010, p. 399–406.
- Monga V, Li Y, Eldar YC. Algorithm unrolling: interpretable, efficient deep learning for signal and image processing. *IEEE Signal Process Mag*. 2021;38(2):18–44.
- Hammernik K, Klatzer T, Kobler E, Recht MP, Sodickson DK, Pock T, Knoll F. Learning a variational network for reconstruction of accelerated MRI data. *Magn Reson Med*. 2018;79(6):3055–71.
- Sun J, Li H, Xu Z et al. Deep ADMM-Net for compressive sensing MRI. *Adv Neural Inf Proces Syst*, 2016;29:10–18.
- Eldar YC, Kutyniok G. Compressed sensing: theory and applications. Cambridge: Cambridge University Press; 2012.
- Unser M. A representer theorem for deep neural networks. *J Mach Learn Res*. 2019;20(110):1–30.
- Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the Brain. *Psychol Rev*. 1958;65(6):386.
- Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature*. 1986;323(6088):533–6.
- Cybenko G. Approximation by superpositions of a sigmoidal function. *Math Control Signals Syst*. 1989;2(4):303–14.
- Telgarsky M. Benefits of depth in neural networks. In: Conference on learning theory. PMLR; 2016, pp. 1517–1539.
- Eldan R, Shamir O. The power of depth for feedforward neural networks. In: Conference on learning theory. PMLR; 2016, pp. 907–940.
- Raghu M, Poole B, Kleinberg J, Ganguli S, Sohl-Dickstein J. On the expressive power of deep neural

- networks. In: International conference on machine learning. PMLR; 2017. pp. 2847–2854.
19. Yarotsky D. Error bounds for approximations with deep ReLU networks. *Neural Netw.* 2017;94:103–14.
 20. Bishop CM. Pattern recognition and machine learning. New York: Springer; 2006.
 21. Schölkopf B, Smola AJ, Bach F, et al. Learning with kernels: support vector machines, regularization, optimization, and beyond. London: MIT Press; 2002.
 22. Vapnik V. The nature of statistical learning theory. New York: Springer Science & Business Media; 2013.
 23. Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings; 2011, pp. 315–323.
 24. LeCun YA, Bottou L, Orr GB, Müller K. Efficient BackProp. In: Neural networks: tricks of the trade, Lecture notes in computer science. Berlin, Heidelberg: Springer; 2012. p. 9–48.
 25. Fukushima K. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol Cybern.* 1980;36(4):193–202.
 26. Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol.* 1962;160(1):106–54.
 27. Qayyum A, Anwar SM, Awais M, Majid M. Medical image retrieval using deep convolutional neural network. *Neurocomputing.* 2017;266:8–20.
 28. Ibtehaz N, Rahman MS. MultiResUNet: rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Netw.* 2019;121:74–87.
 29. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* 1989;1(4):541–51.
 30. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: IEEE conference on computer vision and pattern recognition, 2015, p. 3431–3440.
 31. Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics; 2015. p. 315–323.
 32. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res.* 2014;15(1):1929–58.
 33. Hosseini SAH, Yaman B, Moeller S, Hong M, Akçakaya M. Dense recurrent neural networks for accelerated MRI: history-cognizant unrolling of optimization algorithms. *IEEE J Select Topics Signal Process.* 2020;14(6):1280–91.
 34. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge: MIT Press; 2016.
 35. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput.* 1997;9(8):1735–80.
 36. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. In: NIPS 2014 Workshop on Deep Learning, December 2014.
 37. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Adv Neural Inform Process Syst. 2017, pp. 5998–6008.
 38. Cha E, Oh G, Ye JC. Geometric approaches to increase the expressivity of deep neural networks for MR reconstruction. *IEEE J Select Topic Signal Process.* 2020;14(6):1292–305.
 39. Zhou D-X. Universality of deep convolutional neural networks. *Appl Comput Harmon Anal.* 2020;48(2):787–94.
 40. Belkin M, Hsu D, Ma S, Mandal S. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proc Natl Acad Sci.* 2019;116(32):15849–54.
 41. Belkin M, Hsu D, Xu J. Two models of double descent for weak features. *SIAM J Math Data Sci.* 2020;2(4):1167–80.
 42. Chen SS, Donoho DL, Saunders MA. Atomic decomposition by basis pursuit. *SIAM Rev.* 2001;43(1):129–59.
 43. Daubechies I, Defrise M, De Mol C. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun Pure Appl Math.* 2004;57(11):1413–57.
 44. Xin B, Wang Y, Gao W, Wipf D, Wang B. Maximal sparsity with deep networks? *Adv Neural Inf Process Syst.* 2016;29:4340–4348.
 45. Chen X, Liu J, Wang Z, Yin W. Theoretical linear convergence of unfolded ISTA and its practical weights and thresholds. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18). Red Hook, NY, USA: Curran Associates Inc.; 2018, pp. 9079–9089.
 46. Liu J, Chen X, Wang Z, Yin W ALISTA: analytic weights are as good as learned weights in LISTA. In: International conference on learning representations. 2019.
 47. Solomon O, Eldar YC, Mutzafi M, Segev M. SPARCOM: sparsity based super-resolution correlation microscopy. *SIAM J Imag Sci.* 2019;12(1):392–419.
 48. Wang Z, Liu D, Yang J, Han W, Huang T. Deep networks for image super-resolution with sparse prior. In: Proceedings of the IEEE international conference on computer vision. 2015, pp. 370–378.
 49. Hauptmann A, Lucka F, Betcke M, Huynh N, Adler J, Cox B, Beard P, Ourselin S, Arridge S. Model-based learning for accelerated, limited-view 3-D photo-acoustic tomography. *IEEE Trans Med Imaging.* 2018;37(6):1382–93.
 50. Adler J, Öktem O. Learned primal-dual reconstruction. *IEEE Trans Med Imaging.* 2018;37(6):1322–32.
 51. Solomon O, Cohen R, Zhang Y, Yang Y, He Q, Luo J, van Sloun RJG, Eldar YC. Deep unfolded robust PCA with application to clutter suppression in ultrasound. *IEEE Trans Med Imaging.* 2020;39(4):1051–63.
 52. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE

- conference on computer vision and pattern recognition. 2016, pp. 770–778.
- 53. Dardikman-Yoffe G, Eldar YC. Learned SPARCOM: unfolded deep super-resolution microscopy. *Opt Express*. 2020;28(19):27736–63.
 - 54. Hubel DH, Wiesel TN. Receptive fields of single neurones in the cat's striate cortex. *J Physiol*. 1959;148(3):574–91.
 - 55. Quiroga RQ, Reddy L, Kreiman G, Koch C, Fried I. Invariant visual representation by single neurons in the human brain. *Nature*. 2005;435(7045):1102–7.
 - 56. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. *Science*. 2000;290 (5500):2323–6.
 - 57. Tenenbaum JB, De Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. *Science*. 2000;290(5500):2319–23.
 - 58. Kingma DP, Welling M. Auto-encoding variational Bayes. In: International conference on learning representations. 2014.
 - 59. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. *Adv Neural Inform Process Syst*. 2014, 27.
 - 60. Rezende D, Mohamed S. Variational inference with normalizing flows. In: International conference on machine learning. PMLR; 2015, pp. 1530–1538.

Part II



Introductory Approaches for Applying Artificial Intelligence in Clinical Medicine

4

Niklas Lidströmer, Federica Aresu, and Hutan Ashrafiyan

Contents

Introduction	58
Short AIM History	59
Electronic Health Records (EHRs) and AIM Interaction	61
Learning Health Systems	62
Some Industrial Cases – An Overview	63
AIM Applications	64
Financial Aspects of AIM	65
Emerging Markets	66
Clinical Decision Support Systems	67
Conclusion	71
References	71

N. Lidströmer (✉)

Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden

e-mail: niklas.lidstromer@ki.se; niklas@lidstromer.com

F. Aresu

KTH Royal Institute of Technology, Stockholm, Sweden
e-mail: aresu@kth.se

H. Ashrafiyan

Department of Surgery and Cancer, Imperial College London NHS Trust, London, UK

Institute of Global Health Innovation, Imperial College London, London, UK

Hamlyn Centre for Robotics and Artificial Intelligence,
Department of Surgery and Cancer, Imperial College London, London, UK

e-mail: h.ashrafiyan@imperial.ac.uk

Abstract

The urge of computerized, automatized medical decision making as well as having more efficient and organized health data records for financial and medical purposes have brought the necessity to introduce artificial intelligence algorithms to healthcare.

The first artificial intelligence applications in the medical field were to be seen in the introduction of Electronic Health Records followed by the development of Learning Health Systems and Clinical Decision Support systems.

Currently, the development and increment of artificial intelligence applications by larger and smaller entities from all over the world is in an ongoing process, following the market and its needs.

Keywords

Artificial Intelligence · Medicine · Basic Concepts · Introduction · Classification · Healthcare · Electronic Health Records · Clinical Decision · Support Systems · Learning Health Systems · Medical Education · Machine Learning

Introduction

Artificial intelligence in medicine (here abbreviated AIM) encompasses the use of complex algorithms and software to mimic or estimate the human cognition and analysis of complex medical and health-related data. The definition of AI in healthcare is fluidly changing with progress.

Before attempting to properly define AIM, it is noteworthy to remember the concept's "inborn autophagous definition problem." With this we refer to the fact that AIM users tend to forget the progress as they get used to them. For instance, the auto-interpretation of ECG, an early example of AIM, is often not referred to as AIM by many users. We humans very quickly adapt and take innovations for granted, hence as AIM advances, its end-users tend to think that it "can't be counted as AI," and start to expect something more advanced in the future.

It should be pointed out that, in this textbook on AIM, we refer to *narrow AI* (see definitions of AI in this section), i.e., very well-defined tasks, where AI can outperform humans, and not to *general AI*.

"**AI in medicine** refers to the use of **artificial intelligence** technology/automated processes in the diagnosis and treatment of patients who require care."

"**Artificial Intelligence in Medicine** mainly **uses** computer techniques to perform clinical diagnoses and suggest treatments. AI has the capability of detecting meaningful relationships

in a data set and has been widely **used** in many clinical situations to diagnose, treat, and predict the results."

Here are the early **examples of artificial intelligence in healthcare**.

- **Artificial Intelligence** assistance for general well-being purposes, such as fitness etc.
- **AI-assisted robotic surgery.**
- Clinical judgment or diagnosis.
- Precision medicine.
- Drug discovery (pharmacovigilance).
- Personalized healthcare.

At the core of these procedures is machine learning. Its algorithmic approach is designed to recognize behavioral patterns and from these produce deductive reasoning. Before algorithms can be used further, they must be tested and optimized, in their architectural design and hyperparameters, numerous times to diminish the error margins.

The algorithms in AI naturally differ in numerous ways from humans, but to just set out two main paradigms of major difference:

1. Algorithms read literally – they follow the goal set and obey explicitly according to algorithm layout.
2. Algorithms are scientifically black box objects (can only be viewed in terms of their inputs and outputs) – they can predict and generate data without having previous knowledge about causality and relationship between input and output data [1]. However, released in June 2020, an autoregressive deep learning model by OpenAI, called Generative Pre-trained Transformer 3 (GPT-3) has been created that is able to generate human-like text data by pre-analyses of huge amounts of input text data for their meaning and their relationship with other words [2].

The underlying algorithm in AIM must aim at the analysis of the logical connection between the seen clinical results in a patient and what treatment or prevention is deemed causative (Fig. 1). In other words, aiming at the relationship between

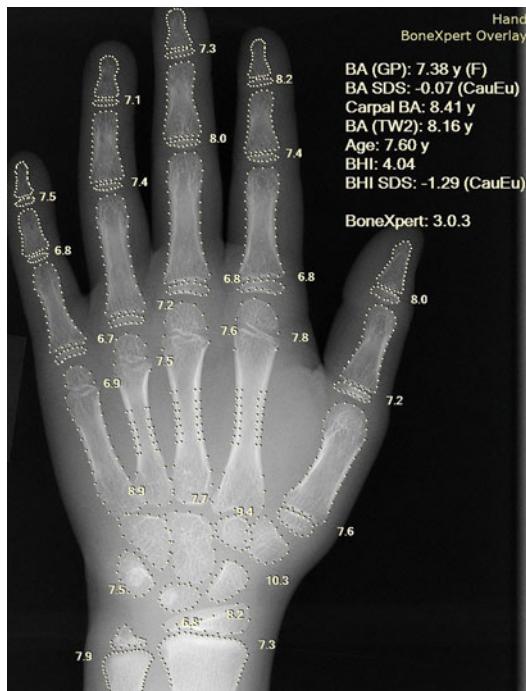


Fig. 1 A casual left-hand X-ray, with automated calculation of the relative bone age, using AIM powered software, BoneXpert. Taken on a Caucasian male 16.27 years old, with delayed puberty. The bone shows the age of 13.49 years by the Gruenlich and Pyle algorithm and 13.21 by the Tanner-Whitehouse algorithm. With these values a final adult height of $188.3+/-3.6$ cm can be approximated [3]

an input (treatment, prevention, procedure, diagnostics) and the output (the result in patient or diagnostic conclusion) [4].

With these above capabilities AI applications are used to ease, monitor, analyze, or respond to treatment regimens, personalized healthcare, diagnostic procedures and even the de novo development of pharmaceutical substances and chemical screening to find lead candidates. This former screening can be done much more efficiently than any chemist could obtain, and hence they could focus on more complex tasks, once released from mere routine work in labs.

Several of the world's leading clinics and healthcare systems such as Massachusetts General Hospital, the National Health Service in the UK, The Mayo Clinic, and the Memorial Sloan Kettering Cancer Centre have contributed to the

development of tools based on AI algorithms for their respective needs. In most cases the development has been possible through collaboration with large tech corporations, or with prominent newcomers such as the software companies Aysadi and Welltok [5].

With these AI-powered systems, hospitals can, apart from the obvious potential medical benefits to patients and their families, also economize with resources, improve the hospital working environment, and manage customer relations. It is necessary, in, e.g., a political context, to emphasize the monetary and purely operational benefits of AIM [6]. It is maybe a laconic conclusion that medical advantage per se will not be sufficient to convince hospital managements or politicians, but needless to say the use of predictive analysis in use by the latter, which optimize the resource usage and other factors [7].

The long-term AIM effects on the health industry, and outside of the medical realm, remain to this date held in obscurity. An in-depth investigatory survey was carried out among the economic faculty specialists, especially economists, with the query of if and how AI, in general, and robotics and machine learning applied into a broad multitude of civic fields, would affect long-term unemployment. The consensus turned out to be a considered net benefit – under the condition's productivity gains were (oddly or evenly) redistributed [8].

Short AIM History

It can be argued at what time period the historical exposé of AI per se should be commenced in a review, since the concept of AI is like a snake eating its own tail – as soon as something exists commonplace and is deemed too simplistic, it is continuously redefined as being outside of AI. A speedometer in a car is definitely far from it, lane keeping assistance used to be high tech, like also thermostats and other trivial items, today taken for granted.

The technique behind this early AI was based on 1960–1970s research, which also gave rise to the Dendral, then considered a problem solution software, i.e., a software designed to emulate the

decision of a human expert [9]. The fields in which this software operated was organic chemistry, leading up to the MYCIN system, which is often regarded a very important major step in the then nascent AIM applications [10, 11], and which evolved into more advanced systems in the 1980s, but the use was limited and never reached any breakthrough in clinical medicine. Maybe the explanation is one of the etiological components of the AI winter that followed [12].

During this following period up until the 1990s the microprocessor seriously emerged on the market, providing the hardware prerequisite for further AIM advancements. AIM must deal with the presence of imperfect input, given the nature of medical science in theory and the patient and doctor inputs, i.e., clinical case inputs, in practical medicine. To then build AIM on medical expertise, would rather lead to the emergence of medical diagnostic decision support systems [13].

To tackle the special criteria in healthcare, the AI programming grounds must be based on, e.g., artificial neural networks (e.g., in cardiology [14] or in oncology [15]), Bayesian networks [16], and the fuzzy set theory (in mathematics aka uncertain sets, of what are somewhat like sets, whose elements have merely degrees of unifying membership) [17]. It is archetypical that these advanced mathematics must be applied for the nature of healthcare AI, with its abundance of previous misinterprets, vaguely presented patient verbalization of symptoms and other patterns, which cannot be identified with AI lacking deeper components of machine learning.

Other vital components in the AIM evolution would hence include milestones, such as the computer hardware per se, augmenting the processor power, enabling rapid collections and handling [18]. The AIM evolution is parallel with the development of graphics processing units (GPUs), which have unlocked new possibilities in gaming, content creation, machine learning, and more. The GPU evolves as a massively parallel component to its close cousin, the central processing unit (CPU). The digitalization in itself of the health records, with widespread EHRs, lead to increased volumes to feed and test the AIM algorithms [19]. Simultaneously we saw the emergence of

robot-assisted surgery and its enhanced precision [20].

Furthermore, large genetic international projects, such as the HUGO project, with human genome mapping, have created an exponential growth of the genomic sequencing databases [21].

Likewise, the outburst of EHR usage worldwide has, as mentioned, led to an increase in the actual substrate to the AI algorithms, by providing big data for the algorithmic AIM foundation [22]. The growth of the EHR contents will likely not decrease with the introduction of computer vision [23] and natural language processing [24], both important AIM achievements, aiming at human perception mimicking.

In 2018 Google launched its DeepMind artificial intelligence system to work with recognition of eye diseases (Fig. 2). The DeepMind system made correct diagnoses 94.5% of the time in a trial with Moorfields Eye Hospital [25].

Examples of AI applications in Medicine that initially brought high benefits to the respective medical specialties are:

- *Preoperative imaging:* For instance, in orthopedic surgery to more exactly plan and evaluate the operative procedures, as well as in complex maxillofacial surgery of several categories, e.g., the probable aesthetic outcomes can be calculated with AIM [27].
- *Diagnostic imaging and radiology:* The most radical soon coming changes are probably to expect within this field [28]. Several radiological societies' [29] annual symposia regularly implement AIM imaging discussions and novel study presentations. In these discussions, the fears [30] and pros [31] of the game-changing AIM applications are ventilated with different weighting toward advantage or disadvantage to the specialty.

Likely all routine analysis of diagnostic imaging will be taken over by AIM. The radiologist will only focus on the complex cases, to check the outcomes, and discuss with colleagues and patients directly, the latter hitherto unusual. The AIM will likely aid in delivering, e.g., a more refined and detailed X-ray result – a human doctor

Fig. 2 Google DeepMind used in eye disease recognition [26]



does occasionally miss some minor X-rays details. For instance, a Stanford study programmed an AIM algorithm used in radiology, which could diagnose pneumonia in a given pulmonary site with a higher precision, i.e., higher mean F1 score (more accurate) than the human radiology specialists included in the study [32].

For several clinical safety reasons, the temporarily novel use of telemedicine with, e.g., mobile applications, and the steeply rising costs for society, tax payers, or insurances, have raised serious concerns among medical professionals and researchers alike – the diagnostic lack of precision via an app has been appalling and likewise the subsequent treatment, often without any decent follow-up. Several reviews in, e.g., the Scandinavian countries have raised serious concerns about the professionalism in using online applications to diagnose and treat patients in these manners, and moreover the spiral of costs have been beyond control at present (2018–2019). In theory the parallel rise of plausibly easing AIM online apps, or the powering of the mentioned doctor-online services with such AIM applications, could at least momentarily tilt this imbalance [33]. Such apps could facilitate patient monitoring at distance and improve the interface to boost the

intractability – a suggestible approach could be synchronized items of the internet of things, such as those wearable, inoperable, or in any other way with measurable assets. In any case, gadgets with continuous monitoring will surely be used by clinic or hospital bound healthcare, and could in real time update the EHRs with uninterrupted flow of data points. Such handling, especially combined with AIM, would notice even minute changes, maybe invisible to the human eye and habitude.

Electronic Health Records (EHRs) and AIM Interaction

EHRs were initially developed and implemented for billing purposes and workflow management; meanwhile few prototypes of EHRs were used for sharing information between physicians as well as between healthcare providers and organizations.

Due to poor design, user unfriendliness, combined with the presence of inefficient workflow ergonomics on healthcare structures, EHRs are deemed to cause burnout in healthcare professionals [34].

As a consequence of EHRs' mismanagement, many healthcare professionals are forced to spend

a large proportion of their working day instead as data input engineers.

However, they still represent the crucial step of digitalization of the healthcare industry and clinics, and their information spread. Some partial relief, or only input technique, could be the use of natural language processing (NLP) tools. Several other EHR automation projects are underway. The adequate integration of AIM into the EHRs will be a crucial step for diagnosis prediction, or mere decision support in the future (Fig. 3).

EHRs are treated in a full chapter in the second part of this book, and especially considering their central role in AIM, if synchronized wisely with clinical decision support systems (CDSSs), automated radiographic interpretations, and other relevant tools and applications.

Learning Health Systems

In deep connection with two of the most frequently used terms in this book, AIM and deep medicine, lie the vital concept of **Learning Health Systems** (LHS) [36].

Fig. 3 Electronic Health Record (EHR) [35]



All in all, the LHSs are EHRs where auto-educational and hence self-improving processes are fused with the medical record system [37] – i.e., on a daily working basis in a clinic the EHRs would become better the more you use them [38].

One early idealistic idea for LHSs was to let evidence-based medicine (EBM) become the scaffolding, or role model for under what conditions or rules the LHSs would be auto-learning [37].

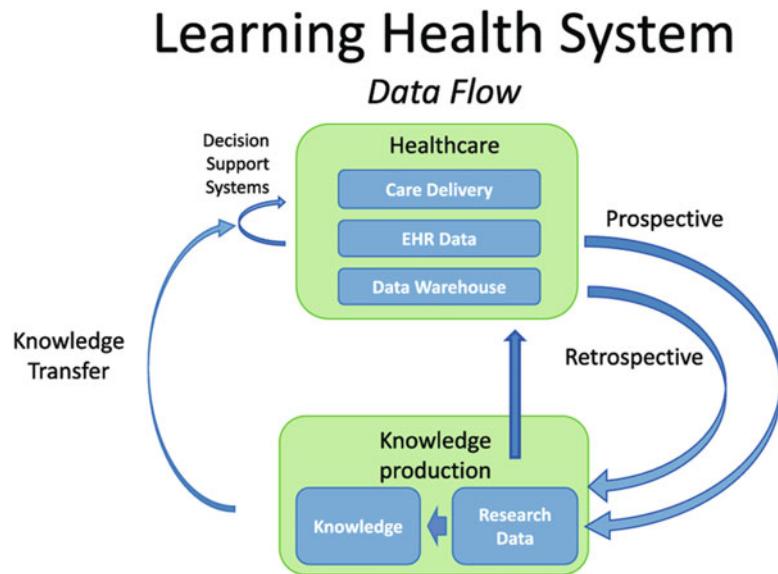
It is also worth noting that not all clinics, or very few, follow the sharpest EBM principles; rather their experience or clinical-practice-based evidence and the differences can be great [39].

The growing field of AI powered EHRs have led to the foundation of the medical journal *Learning Health Systems* from Wiley Publishers [38] (Fig. 4).

The LHS usage is based on the EHR and receives big data substrate from it [41].

The LHSs have, over the last few years, evolved into containing the following core components [42]:

Fig. 4 A model of Learning Health Systems [40]



1. The platform for possible research even at local clinic level
2. Personalized statistic modules for each individual patient
3. Synchronization of clinicians and researchers to handle the growing mass contents
4. Structured use of big data
5. Organization supporting subgroups and communities of patients or professionals
6. Intercommunication tools

The natural circuit of an LHS encompasses medical data collection from a large data set of patient cases, and a problem formulation derives [43]. This data compilation and its inputs is the fruit of work carried out by healthcare professionals. With correct use of computer programs this substrate can be analysed, the assembled science may affect future treatment practices, i.e. improved algorithms - and when the data quality per se is augmented, an auto-enhancing substrate cycle arise [45].

The following taxonomy can be used for LHS classification subtypes:

- (i) Assortment with cohort identification, that is, patients with similar patterns [46]

- (ii) Care optimization benchmarking – searching for better treatment options – with positive deviance [47]
- (iii) Suboptimal care identified with negative deviation [44]
- (iv) High-risk patient identification – with pattern recognition [48]
- (v) Identification of high-risk healthcare situations with suboptimal care (predictive risk)
- (vi) Algorithms applied to clinical patient data evolve into treatment recommendations
- (vii) The efficacy of different treatment regimens are compared and the best one identified
- (viii) With AI medical routine work, e.g., diagnostic image interpretation, automated
- (ix) Monitoring of epidemics, adverse effects, misconduct, or other clinical issues [49]

Some Industrial Cases – An Overview

There is power in numbers, and in big data big is better. When large companies merge their data synergistic effects arise, allowing access [50] and substrate feed for more precise AIM, more correctly adjusted machine learning algorithms

with higher generalization capabilities. The following is a short introductory overview of companies engaged in these processes – some of the business cases coming out of these and other companies, along with bright start-ups, will be further elaborated later.

A major AIM focus has hitherto been on Clinical Decision Support (CDS) systems, and with increased big data, the CDS precision rises [51].

Some major corporations with AIM projects are to this date:

- **Google:** *The DeepMind* platform is currently being used by the NHS (National Health Service) in the UK. The division of DeepMind Health contributes to the improvement of public healthcare with, for instance, a project on risk-detection for chronic diseases via the patients' mobile app, and another project with computer vision algorithms to detect tumor cells with the potential to evolve into cancer.
- **Intel:** Acquired *Lumiata*, which uses AI algorithms to help targeting risk patients and present different treatment alternatives [52].
- **IBM:** The broadly marketed *Watson* project, to be discussed further later. Now the branch *Watson Oncology* is under development together with the Cleveland Clinic and the Memorial Sloan Kettering Cancer Centre [53]. IBM has also worked tightly with Johnson & Johnson on scientific paper analysis, aiming to process this further into AI-directed drug innovation. CVS Health is looking into AI mobile apps for the personal management of chronic medical conditions [54].
- **Microsoft:** With Oregon Health & Science University (the Knight Cancer Institute), Microsoft has initiated the *Hanover Project* to analyze oncological research and thus obtain the optimal cancer treatments [55]. They are also involved in a project on programmable cells and image processing of tumor progression and prediction of tumoral cell action [56].

Some healthtech start-ups worth mentioning are focusing their work on automatizing prescribed medication administration and distribution,

as well as fully automatizing diabetic retinopathy and mammary gland tumor detection, including real-time user's data in EHRs from a multitude of sources and helping blind people to navigate using pathway recognition algorithms.

AIM Applications

There is of course a plethora, expanding at an exponential pace, of others developing many other AIM-driven applications, such as chatbots to schedule patient bookings in the EHRs and to optimize clinic resources, among the top-ten rated medical AI-driven applications with the so-called personal assistance at present [57].

AIM application entrepreneurs use several archetypical manners to gain and maintain market access and their common modus operandi has become an economic research field in itself, with revealing patterns. Their business plans fall into some typical categories and depend on the target user value (patient, healthcare providers, or other payer) and the mechanisms used to capture the value per se (to provide valuable information, the product of Big Data mining, or by connecting groups of interest, to mention a few) [58].

Government-owned companies and private entities launched mobile apps that directly or indirectly focus on healthcare using artificial intelligence approaches.

Among currently widespread important apps for mobiles can be mentioned the free *Symptomate* (from Infermedica), which is a symptom-checker on Google play. The same company has launched a personal AI voice guide, to respond to a person's symptoms, and made available for *Alexa* (Amazon), *Google Assistant*, and *Cortana* (Microsoft).

The instep level of the applications' output depends on whether aiming toward medical professionals or laypeople. In case of just common medical knowledge, there are at least three major, rapidly growing medical apps – Your MD, GP at Hand (from Babylon), and Ada Health. These pretend to mimic a medical consultation to some extent. The AIM fundamentals of these apps are in

all three only what the individual types in a previous medical history. The reports of his/her history in verbal form and voice recognition classifies the input and runs it against a database of all plausible diagnoses and adds suggested actions, respectively [59].

There are also a couple of apps monitoring delirium and anxiety levels in admitted patients, especially in seniors, impaired patients, or those suffering from different types of dementia. These patient subgroups are deemed to especially benefit from an adapted assisting interface, where extra attention is paid to act as if equipped with human EQ [60].

Economic studies [58] in 2019 have concluded medical expenditures will decrease when AI-powered diagnostics with more accurate prognosis allows planning of resources in a proper manner.

The healthcare effectiveness will likely increase significantly with the use of Virtual nursing assistants – the equivalences are already in use in AI-driven customer service of many companies. With voice recognition and AI, “synthetic nurses” over the phone could help patients with, e.g., mass-triage and other routine phone checks. Fears have been raised that AI would replace healthcare jobs, but most debaters agree AI would rather stay as a decision support system, streamline routine work, and allow healthcare persons to spend more time bedside and in human interaction.

Financial Aspects of AIM

The AI health market has shown explosive growth. Acquisitions of AI start-ups are rapidly increasing while the health AI market is set to register an explosive Compound Annual Growth Rate (CAGR) of 40% through 2021. The total AI health market size was estimated to \$600 M in 2014 and to \$6.6B in 2021 [61]. Artificial Intelligence (AI) in the healthcare market is expected to reach the market value of US\$ 26.6 billion by 2025, growing at a CAGR of 41% during the forecast period (2018–2025). AI in the healthcare market is estimated to grow with 15.7 trillion USD until 2030 [62].

AI is rapidly entering the healthcare sector, especially growing in areas such as radiology and cancer detection. It is already there in many places and is poised to take over many tasks. This in turn of course raises the question of whether it would replace nurses and physicians, and under which economic implications.

As will be shown, a wide range of AI finance experts deem these jobs are definitely not in danger. That said, both AI and machine learning are in a prime position to alter clinical workflows and the training of MDs. And with the market growing in the currently explosive manner, the AI healthcare implementation is inevitable [63].

The human hands and senses are definitely needed, not only for the data input of gigantic amounts of facts, but also to analyze, summarize, conclude, and holistically communicate also in a human emotional manner. And AIM is now utilized mostly to organize and aggregate data – looking for trends and patterns and making recommendations.

To make a brief expert review, the PeriGen CEO Matthew Sappern puts no stock in the theory that clinicians’ jobs are in jeopardy. Instead, he looks at AI more as an empowerment tool: “I think it does things that are really imperative that are not necessarily what nurses can do,” he says. “These tools are not so great where reasoning and empathy are required. You teach them to do something, and they will do it over and over and over again, period. They’re good tools to provide perspective, but it’s all about the provider or nurse who’s making sense of that information.”

He also believes that “AI can help nurses focus more on the actual job of nursing, and focus more on the abstract things that can truly impact patient care. And it has the potential to increase their confidence, as they can report back to the doctor with hard stats instead of vagaries. Used wisely and it can be a boon to fact-based clinical observation.”

The Chief Product Officer of Jvion, Dr John Showalter, does agree with Matthew Sappern. He stresses that it is the hype in itself about jeopardized jobs which is so scary, but that the reality is

not. He states that “there are absolutely places where AI is ready to go today, and then there’s a whole bunch of AI hype that’s really scary, so sorting out the AI that’s ready to help patients and the hype can be really difficult for leadership.”

As demonstrated in the article “AI doctors and engineers are coming – but they won’t be stealing high-skill jobs,” [64] the above quoted experts’ views are shared with a broad range of key opinion leaders (KOLs). The focus of this chapter is rather the plethora of empowering AI applications, which will only be beneficial to patients and clinicians.

Emerging Markets

Rapidly growing markets in formerly called developing nations could, as in many other areas, jump several steps in the technical evolution, and right into the age of AI in healthcare. The care becomes available worldwide in a new manner with new AI applications, coming in the new emerging markets, according to several investment agencies [65].

Radiological interpretations are mentioned several times in this book – here the use of AI empowers diagnostic imaging labs to diagnose many more cases with increased speed 24/7 all year round, and hence we can also use all expensive heavy equipment much more effectively. The AI can also help to train specialist radiologists and other medical professionals, but it is always estimated human intervention will be needed in more complex cases and where patient–doctor interaction is required.

For instance, in 2020, the CAD4TB was developed to detect tuberculosis-related abnormalities in ordinary anterior-posterior chest X-rays. This computer-aided detection software takes a single chest X-ray as its input, in the form of a DICOM image (Digital Imaging and Communications in Medicine, the standard for the communication and management of medical imaging information and related data), and produces several outputs: a quality assessment of

the input image, a heat map highlighting possible abnormal areas, and a score between 0 and 100 indicating the likelihood of the X-ray being abnormal and the subject on the X-ray being affected by tuberculosis. This deep learning-based software is now installed in over 150 mobile systems in 30 countries [66].

Today there are many successful applications of deep learning in medical imaging. An author review is presented recently [67].

The CAD4TB initiator Professor Bram van Ginneken, also co-pioneered the *Grand Challenge* platform for end-to-end development of machine learning solutions in biomedical imaging. It provides the public datasets and the software solutions of the winners of the international “challenge competitions.” At the time of the 1st edition of this book there are outcomes and data from 238 challenges [68].

In statistics and machine learning the concept of “ground truth” is a term relative to the knowledge of the truth in a specific question. To reach this level is the ideal [69]. The term is used in statistical models to strengthen or weaken research hypothesis. To gather objective data, which can be proven, is also called *ground truthing*. The term can be compared with *gold standard*, which in statistics and medicine refers to a standard test; a diagnostic test, or benchmark, which is deemed to be the best [70].

Before any broad implementation of AIM, it’s necessary to test and validate the algorithms behind the applications. Algorithmic biases, machine morals, and DNR (do not resuscitate) conditions, will be elaborated further in the AIM ethics chapter. These mentioned factors have lain behind the early regulatory attempts within the AIM fields [71]. In most countries there is currently almost no regulation legislation within the fields of AIM.

There have been a couple of legislatively preparatory workshops held in some countries during the last 3 years, 2016–2019. The new EU legislation GDPR does though indirectly affect the AI field, but in the EU, and elsewhere, the legislation that affects the AIM field, needs to undergo rapid adaptation to fit the ethical issues.

Clinical Decision Support Systems

The term **Clinical Decision Support System (CDSS)** includes some of the above, and is designed to assist the physician with medical decision support, in especially very complex clinical cases. In this way these systems may provide a bridge between clinical observations with medical science and have an impact, depending on background algorithms, to affect the ultimate choices made by medical doctors in a sharp medical setting. The CDSS is of central interest for the AIM construction (Fig. 5).

In the book the very authentic motif of EHR introductions were discussed, and in the same manner the actual need and effectiveness of a CDSS have been debated intensely. Could the usage of CDSSs and their effectiveness be considered evidence based?

When a CDSS and an EHR was combined, which was studied in a 2014 review, some benefits were maybe shown; however there was no effect on survival at all [73].

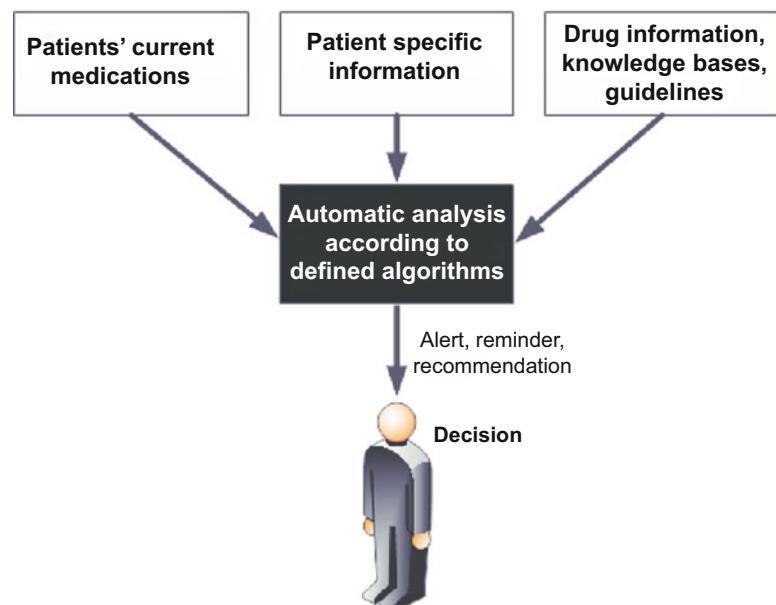
One of the earliest systematic reviews on this topic in 2005 writes in its conclusion that 64% of the included studies showed improvements of the medical care giver, and moreover patient outcomes were improved in 13% of the studies in

the review. Some key features that were connected to this improvement include e-alerts inside the system without the user having to actively switch on this interface. This study, which was published in the JAMA, also concluded that overall the studies (1973–2004), included in the review, show an improved methodological quality in the last part of this period and the number of studies also steeply increased in number [74].

The very same year another review in the BMJ showed a 68% improvement in its included body of studies. The authors did also pinpoint the key ingredients of a successful CDSS; it must be integrated in the clinical work altogether, be it electronic and not paper based, the decision support must be given on time, that is when the patient visits the clinic, in the very consultation and not before or after, and finally it must provide treatment recommendations [75].

The picture is however not unanimous. Less positive reviews have been published as well, e.g., in 2011, a study demonstrated that there was a significant discrepancy between the benefits which were stated about CDSSs and other similar applications, and what could actually be shown empirically. The lack of cost-effectiveness was not demonstrated as they concluded in their review. [76].

Fig. 5 Principle of a Clinical Decision Support System (CDSS) [72]



Until 2014 there was likely no real long-term evaluation to be found in peer-reviewed journals. But this year, probably the first 5-year follow-up of the CDSS's clinical effectiveness was published, and studied the long-term effect of computer-assisted decision support for antibiotic treatment in critically ill patients – as a prospective “before/after” cohort study [77]. It concluded that the implementation of computerized regional adapted guidelines for antibiotic therapy is paralleled with improved adherence. Even without further measures, adherence stayed high for a longer period and was paralleled by reduced antibiotic exposure. Improved guideline adherence was associated with reduced ICU mortality.

The CDSS is characterized by *medical knowledge management*, and uses more than two systematized patient data to generate support for medical decisions by the physician, hence it is also defined as an *active* knowledge management system – the more clinical items about a patient the better.

The very purpose of this system is to support the professional’s clinical decision – the clinician interacts with the system to analyze, diagnose, and choose a treatment. It is not a question of whether the system is taking the decision, rather producing a suggestion or reminder not to miss to consider a diagnosis in a complex case, i.e., the clinician uses both the system and own knowledge to evaluate the cases.

The CDSS group falls into two major categories: they are either based on knowledge or not. For instance, DDSS (Diagnosis Decision Support Systems) suggest diagnoses based on data input – the DDSS response may lead to the health professional’s decision to take further diagnostic action to ring in the correct diagnosis [78].

Case-Based Reasoning (CBR), where similar problems were dealt with previously, form the basis for the problem-solving process in a clinical case. A CBR system can be considered a subtype of CDSS [79].

In business, as in medicine, timing is everything. Even inside the medical support systems the actual is crucial. When CDSS (Clinical Decision Support Systems) are classified the timing factor is divided. Doctors would use the CDSS

when they see the patient in most cases. The CDSS are hence either of pre-, per- (or during), or post-diagnosis type. Again, they are either of a (1) preparatory nature, (2) of the presently filtering type to ease the preliminary diagnosis, or (3) To follow up a diagnosis in past tense or, to mine the big data derived from medical data and clinical knowledge and Evidence Based Medicine (EBM) manuals. The latter would be used for predictive purposes. In the debate, some drop radical ideas such as 80% of all clinical work would be done by these systems instead of by human doctors [80]. But the same author does though conclude that “*data-driven healthcare won’t replace physicians entirely, but it will help those receptive to technology perform their jobs better.*” Given the outcomes of the presented studies in this book, this statement seems too radical and deterministic. Of course, it depends on which specialty we are reasoning about, and what we define as “medical work” – pure routine work or reasoning requiring much higher human capabilities. The former debaters have stated, as previously quoted in this book, that “*if a specific doctor’s quality of medical routine work can be replaced by AIM, then it deserves to be replaced.*”

Others, like the editor of the Healthcare Finance Journal, argue that *while AI has already shown promise in automating certain tasks, it doesn’t seem likely that it will replace flesh-and-blood clinicians anytime soon* [63].

The NHS in the UK uses DDSS to triage the calling patients. With this setup an operator with no medical training can suggest a base level likely action to take – within limited variance though – the suggestions could be: *call 112, see your GP or wait*. No studies have been published on whether this system is better than just average common sense, at least none have been found in this review, but in 2018 a study of another decision support system was published, though with a vast difference – it involved medically trained nurses too – for triage management, with a hybrid approach using rule-based reasoning and fuzzy logic.

This study proposed a rule-based decision support system for triage classification and management. As said, a hybrid approach was applied to

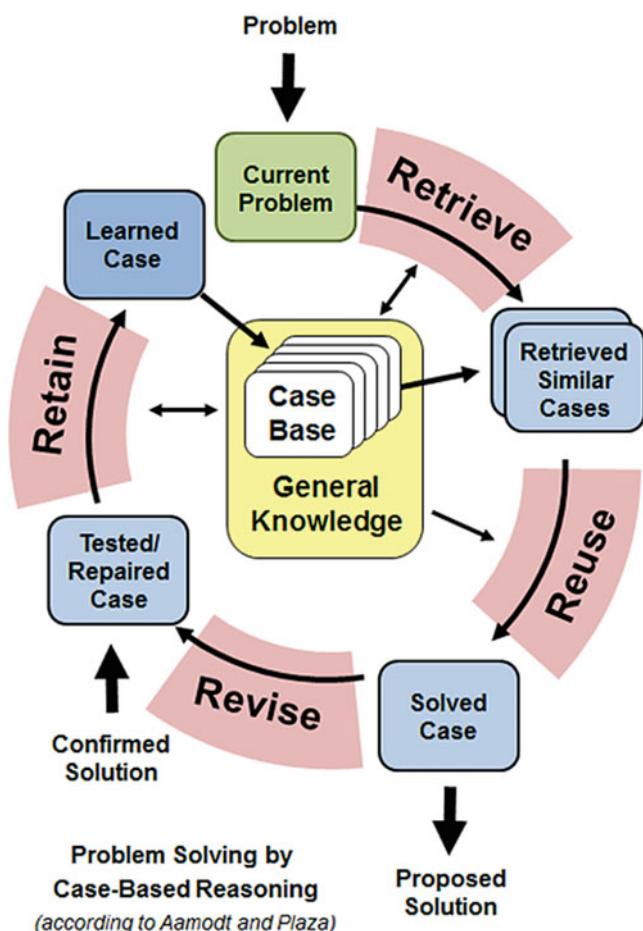
accurate classification of triage based on guideline. Reduction of classification errors and improvement of triage documentation were achieved [82]. It concluded the test-designed system was effective in determining the triage level of patients and it proved helpful for nurses as they made decisions, generated nursing diagnoses based on triage guidelines. The hybrid approach could then reduce triage misdiagnosis in a highly accurate manner and improve the triage outcomes (Fig. 6).

As mentioned above the CDSSs are classified into either having a founding knowledge database or not. If possessing a knowledge base, then the CDSS also contains an *inference engine* and a *communication tool*. In AIM, the inference engine is the part of the system that applies the logical rules governing the information deduction from

the knowledge base. A simple example of such determining law is the type called “IF-THEN rules.” This type of rule is found in action in many EHRs in the form of multidrug interaction checkers – IF drug X plus drug Y, THEN produce alert Z into the user interface of the program. The same core structure has been programmed into the source code of the online WebMD’s multidrug interaction checker, which can handle interactions up to several degrees of complexity, albeit interactions of large numbers of different pharmaceutical medications have not always been studied in detail.

Of course it is of value when the users can update the interaction checker with new drugs or when interactions have been found less valid in further research, or when we’d like to switch off an alert because it is not relevant to the patient –

Fig.6 Basic principle of Case-Based Reasoning (CBR) [81]



compare with the auto-correction in the spell-check function – sometimes the relevant word doesn't exist in the dictionary of the base data collection. Editing and receiving output are both parts of the communication mechanism of the system [78].

To be able to communicate with the computer in a relevant and computable manner, we need to use specific expression languages, e.g., CQL (Clinical Quality Language) or the GELLO. With these, one can set the frames for the system. For instance, we can use these languages to suggest treatment or diagnostic actions, e.g., if a patient has a certain blood pressure of X and Y then the patient should see a doctor within Z weeks. Or if serum potassium level is within a certain range, then a certain action must be taken, etc. The Clinical Quality Language (CQL) is a Health Level Seven International (HL7) authoring language standard that's intended to be human readable. HL7 is a standards-developing organization that provides a framework and standards for the exchange, integration, sharing, and retrieval of electronic health information that supports clinical practice and the management, delivery, and evaluation of health services [83].

The other group of CDSSs is not rule-based, and uses an AI-form termed *machine learning*, which does not require a rule base, since this group's AI structure gets the computer to *learn* from the base of past events (the knowledge base), and/or recognize clinical patterns [84].

The above structure means there is no need for the rules or feedback from medical professionals, as mentioned in the first CDSS group. This latter group is called "black boxes" and act independently, but can of course, evidently, neither reason nor explain its conclusions. This group is only used for cold suggestions and in the vast majority of clinical cases only used after a diagnosis or for pattern recognition and alert human doctors to focus in a certain area, which may require further consideration or research.

The machine learning-based systems, mostly used in clinical decision support systems, include the following two major groups of *support vector* networks, *genetic algorithms* and the use of synthetic or *artificial neural networks* [85].

- **Genetic Algorithms** (GA) contain processes that are evolutionary in their simplified nature – they use a *selection of the fittest solution* [86] to produce the optimal CDSS outcome. As a comparison, in evolutionary cellular biology, the GAs can be understood as a *metaheuristic* that delivers a sufficiently good solution inspired by the process of natural selection that belongs to the larger class of evolutionary algorithms (EA). In computer science, the selection algorithm chooses among randomly occurring problem solutions. The solution picked may be altered, i.e., mutated or recombined, and can be run through the processes again, repetitively, and be refined ultimately. GAs are as mentioned "black boxes," with the goal to extract knowledge from clinical patient data sets [87].
- Artificial Neural Networks (ANN) remind in structure of the group above, but use nodes and weighted inter-node connections and hence discover and analyze the relationship symptom-diagnosis, extracted from clinical patient data sets [78].

The term metaheuristic may require some explanation – it is used in mathematics and computer science and means a procedure (or heuristic) of the higher echelon, designated to produce or identify a search algorithm (heuristic), which can come up with the best solution (the key issue in an optimization problem) [88]. This is relevant especially when the input substrate is imperfect or incomplete, such as in a medical record – so in these circumstances the metaheuristics, which may now be easily understood as general algorithmic frameworks, often nature-inspired, are designed to solve complex optimization problems [89], here within medical data.

At the core of what needs to be solved lies the optimization problem, i.e., to find not only a treatment derived from patient data, but the best treatment, and more so, continuously optimize this superlative over and over. Again, to pick the best solution out of several feasible solutions. Optimization problems are of two natures – the continuous and the discrete – depending on which variables are involved [90].

Yet following guidelines given by clinical decision support systems, more than 30% of patients still do not receive the correct treatment. And in the AIM in EBM (Evidence Based Medicine) we learn that only 60% of patients receive the evidence-based treatment, 30% receive, as just said, an incorrect one, and 10% a directly harmful regimen.

To increase the chances of treating the patients in the best way possible, multimodal clinical decision support systems have been introduced, combining guidelines given by different sources, i.e., clinical practitioners' experience with results [98] given by data mining analysis.

Conclusion

The whole healthcare field largely benefits from the usage of Artificial intelligence, and future applications are expected to determine the efficiency and quality of their services.

AI in medicine will enter literally all areas of healthcare, and in various forms such as image and pattern recognition, decision support systems, robots, applications, drug development, health records, clinical research, medical imaging, etc.

Companies – from start-ups to big-techs – are steeply increasing their investments in medical AI. Profits and applications are fueling new AI tools and applications into new innovative projects, with the goals increased in public and personal healthcare. AI in medicine will boost world economy with 15 trillion USD during the coming 10-year period.

In all medical specialties the effect of AI will be very tangible, and currently the advancements have reached different levels of maturity, and require a plethora of interpretations and approaches to solve their specific needs, which will be elaborated in part III.

In the remainder of part II of this textbook we present the general aspects in common for all areas of AIM – the “lessons for all doctors.” In part III we systematically present AIM in all medical specialties and subspecialties.

References

1. “Algorithms Need Managers Too”, Harvard Business Review. 2016.
2. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Prafulla D, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. Language models are few-shot learners. arXiv:2005.14165. 2020
3. X-ray by courtesy to Dr. Mikael Häggström, radiologist, NU Hospital, Gothenburg, Sweden, 2017. Creator of the WikiJournal of Medicine and Radlines, providing open access guidelines for radiologists.
4. Coiera E. Guide to medical informatics, the internet and telemedicine. London: Chapman & Hall; 1997.
5. Webinar. CB Insights. Artificial Intelligence report, 28 June 2016.
6. Kent J. Providers Embrace Predictive Analytics for Clinical, Financial Benefits. HealthITAnalytics, 08 August 2018.
7. Lee K. Predictive analytics in healthcare helps improve OR utilization. SearchHealthIT.
8. Wallach W. Moral machines, (mentioned in the introductory chapter). Oxford: Oxford University Press; 2010.
9. Lindsay RK, Buchanan BG, Feigenbaum EA, Ledberg J. DENDRAL: a case study of the first expert system for scientific hypothesis formation. Artif Intell. 1993;61(2):209–61.
10. Clancey WJ, Shortliffe EH. Readings in medical artificial intelligence: the first decade. Boston: Addison-Wesley Longman Publishing; 1984.
11. Bruce G, Buchanan BG, Shortliffe ED. Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project. 1984
12. Duda RO, Shortliffe EH. Expert systems research. Science. 1983;220(4594):261–8.
13. Miller RA. Medical diagnostic decision support systems – past, present, and future. J Am Med Inform Assoc. 1994;1(1):8–27.
14. Baxt WG. Use of an artificial neural network for the diagnosis of myocardial infarction. Ann Intern Med. 1991;115(11):843–8.
15. Maclin PS, Dempsey J, Brooks J, Rand J. Using neural networks to diagnose cancer. J Med Syst. 1991;15(1): 11–9.
16. Reggia JA, Peng Y. Modelling diagnostic reasoning: a summary of parsimonious covering theory. Comput Methods Prog Biomed. 1987;25(2):125–34.
17. Adlassnig KP. A fuzzy logical model of computer-assisted medical diagnosis. Methods Inf Med. 1980;19:14.
18. Koomey J, Berard S, Sanchez M, Wong H. Implications of historical trends in the electrical efficiency of computing. IEEE Ann Hist Comput. 2011;33(3):46–54.

19. Dinov ID. Volume and Value of Big Healthcare Data. *J Med Stat Inform.* 2016;4:3. <https://doi.org/10.7243/2053-7662-4-3>
20. "Artificial Intelligence and Machine Learning for Healthcare" Sigmoidal, 21 December 2017.
21. Barnes B, Dupré J. Genomes and what to make of them. Chicago: University of Chicago Press; 2009.
22. Jha AK, DesRoches CM, Campbell EG, Donelan K, Rao SR, Ferris TG, ... Blumenthal D. Use of electronic health records in US hospitals. *N Engl J Med.* 2009;360(16):1628–38.
23. Dougherty G. Digital image processing for medical applications. Cambridge: Cambridge University Press; 2009.
24. Banko M, Brill E. Scaling to a very large corpora for natural language disambiguation. In: Proceedings of the 39th annual meeting on association for computational linguistics. Association for Computational Linguistics; 2001. p. 26–33.
25. Burgess M. "Now DeepMind's AI can spot eye disease just as well as your doctor". Wired UK, 13 August 2018.
26. Credit: Google DeepMind.
27. Patcas R, Bernini DAJ, Volokitin A, Agustsson E, Rothe R, Timofte R. Applying artificial intelligence to assess the impact of orthognathic treatment on facial attractiveness and estimated age. *Int J Oral Maxillofac Surg.* 2019;48(1):77–83.
28. "Artificial Intelligence in Radiology: The Game-Changer on Everyone's Mind". Radiology Business, 13 October 2017.
29. Among others: the Radiological Society of Northern America, European Society of Radiology.
30. Chockley K, Emanuel E. The end of radiology? Three threats to the future practice of radiology. *J Am Coll Radiol.* 2016;13(12):1415–20.
31. Jha S, Topol EJ. Adapting to Artificial Intelligence. *JAMA.* 2016;316(22):2353–4.
32. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz C. CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv:1711.05225.* 2017
33. Pacis D. Trends in telemedicine utilizing artificial intelligence. *AIP Conf Proc.* 2018;1933:040009.
34. Melnick ER, Dyrbye LN, Sinsky CA, Trockel M, West CP, Nedelec L, Tutty MA, Shanafelt T. The Association Between Perceived Electronic Health Record Usability and Professional Burnout Among US Physicians. *Mayo Clin Proc.* 2020;95(3):476–487. <https://doi.org/10.1016/j.mayocp.2019.09.024>
35. "Everything you need to know about electronic health records". Allison Tsai, May 2015.
36. The Learning Healthcare System: Workshop Summary. Olsen L, Aisner D, McGinnis JM, editors. Institute of Medicine (US). ISBN 978-0-309-10300-8, July 2006.
37. Institute of Medicine. Digital infrastructure for the learning health system: the Foundation for Continuous Improvement in Health and Health Care. Washington, DC: Institute of Medicine; 2011. ISBN 0-309-15416-2
38. McLachlan S, Potts HW, Dube K, Buchanan D, Lean S, Gallagher T, Johnson O, Daley B, Marsh W, Fenton N. The Heimdall framework for supporting characterisation of learning health systems. *J Innov Health Inform.* 2018;25(2):77–87.
39. Greene S, Geiger A. A review finds that multicenter studies face substantial challenges but strategies exist to achieve Institutional Review Board approval. *J Clin Epidemiol.* 2006;59(8):784–90.
40. Ethier J-F, McGilchrist M, Barton A, Cloutier A-M, Curcin V, Delaney B, Burgun A. The TRANSFoRM project: experience and lessons learned regarding functional and interoperability requirements to support primary care. *Learn Health Syst.* 2018;2:e10037. <https://doi.org/10.1002/lrh2.10037>.
41. McLachlan S, Dube K, Johnson O, Buchanan D, Potts HW, Gallagher T, Fenton N. A framework for analysing learning health systems: Are we removing the most impactful barriers? *Learn Health Syst.* 2019;3(4):e10189. <https://doi.org/10.1002/lrh2.10189>
42. Forrest C, Margolis P, Seid M, Colletti RB. PEDSnet: how a prototype pediatric learning health system is being expanded into a national network. *Health Aff.* 2014;33(7):1171–7.
43. Taylor P. From patient data to medical knowledge: the principles and practice of health informatics. London: Blackwell Publishing; 2007.
44. Deeny S, Steventon A. Making sense of the shadows: priorities for creating a learning healthcare system based on routinely collected data. *BMJ Qual Saf.* 2015;24:505–15.
45. Abernethy A, Ahmad A, Zafar SY, Wheeler JL, Reese JB, Lyerly HK. Electronic patient-reported data capture as a foundation of rapid learning cancer care. *Med Care.* 2010;48(6):S32–8.
46. Friedman C, Wong A, Blumenthal D. Achieving a nationwide learning health system. *Sci Transl Med.* 2010;2(57):1–3.
47. Bradley E, Curry LA, Ramanadhan S, Rowe L, Nembhard IM, Krumholz HM. Research in action: using positive deviance to improve quality of healthcare. *BMC Implement Sci.* 2009;4:25.
48. Lewis G, Kirkham H, Vaithianathan R. How health systems could avert 'triple fail' events that are harmful, are costly, and result in poor patient satisfaction. *Health Aff.* 2013;32(4):669–76.
49. Ye Y, Wamukoya M, Ezech A, Emina JB, Sankoh O. Health and Demographic Surveillance Systems: a step towards full civil registration and vital statistics in sub-Saharan Africa? *BMC Public Health.* 2012;12:741.
50. Paul MR. What merger mania means for health care. CNN Money, 04 November 2018.
51. Horvitz EJ, Breese JS, Henrion M. Decision theory in expert systems and artificial intelligence. *Int J Approx Reason.* 1988;2(3):247–302.

52. Primack D. Intel Capital Cancels \$1 Billion Portfolio Sale. *Fortune*, 2016.
53. Lorenzetti L. From cancer to consumer tech: a look inside IBM's Watson health strategy. *Fortune*, 5 April 2016.
54. Cohn J. The robot will see you now. *The Atlantic* 2013, 20 February 2013.
55. Knapton S. Microsoft will 'solve' cancer within 10 years by 'reprogramming' diseased cells. *The Telegraph*.
56. Bass D. Microsoft develops AI to help cancer doctors find the right treatments. New York: Bloomberg; 2016.
57. Proffitt C. Top 10 artificially intelligent personal assistants. *Disruptor Daily*, 2017.
58. Garbuio M, Lin N. Artificial intelligence as a growth engine for health care startups: emerging business models. *Calif Manag Rev*. 2019;61(2):59–83.
59. Parkin S. The artificially intelligent doctor will hear you now. *MIT Technology Review* 2016. <https://www.technologyreview.com/2016/03/09/8890/the-artificially-intelligent-doctor-will-hear-you-now/>
60. Haigh L. Bringing AEI technology into hospitals. *International Travel Health Insurance Journal* 2019. <https://www.itij.com/latest/news/bringing-aei-technology-hospitals>
61. Accenture Analysis. accenture.com
62. UnivDatos.com Market Insights, Artificial Intelligence (AI) in Healthcare – Market Size, Trends and Competitive Landscape: Global Market Forecast to Artificial Intelligence (AI) in Healthcare – Market Size, Trends and Competitive Landscape: Global Market Forecast to 20252025. <https://www.prnewswire.com/in/news-releases/artificial-intelligence-ai-in-healthcare-market-to-reach-us-26-5-billion-by-2025-globally-cagr-41-univdatos-market-insights-842269344.html>
63. Lagasse J. Why artificial intelligence won't replace doctors. *Healthcare Finance Journal* 2018. <https://www.healthcarefinancenews.com/news/why-artificial-intelligence-wont-replace-doctors>
64. Minku LL, Levesley J. AI doctors and engineers are coming – but they won't be stealing high-skill jobs. *The Conversation UK*, 20 August 2018.
65. Novatio. 10 Common Applications of Artificial Intelligence in Health Care. Novatio, 30 August 2017.
66. Murphy K, Habib SS, Zaidi SMA, Khowaja S, Khan A, Melendez J, Scholten ET, Amad F, Schalekamp S, Verhagen M, Philipsen RHHM, Meijers A, van Ginneken B. Computer aided detection of tuberculosis on chest radiographs: An evaluation of the CAD4TB v6 system. *Sci Rep*. 2020;10(1):1–11. *Nature.com*
67. Zhou SK, Greenspan H, Davatzikos C, Duncan JS, van Ginneken B, Madabhushi A, Prince JL, Rueckert D, Summers RM. A review of deep learning in medical imaging: image traits, technology trends, case studies with progress highlights, and future promises. *arXiv preprint arXiv:2008.09104*. 2020
68. <http://grand-challenge.org/challenges/>
69. Lemoigne, Yves; Caner, Alessandra (Eds.), Molecular Imaging: Computer Reconstruction and Practice, ISBN 978-1-4020-8750-9. <https://doi.org/10.1007/978-1-4020-8752-3>, Springer Netherlands, 2008.
70. Versi E. "Gold standard" is an appropriate term. *BMJ*. 1992;305(6846):187.
71. Antonia FC, Adam CZ, Alexander RV. Point/Counterpoint: Artificial Intelligence in Healthcare. *Healthcare Transformation* 2017;2(2). <https://doi.org/10.1089/heat.2017.29042.pcp>
72. Hammar, Tora. eMedication – improving medication management using information technology, Linnaeus University Dissertations; 188/2014, ISBN: 978-91-87925-15-3, ORCID iD: 0000-0003-1549-2469, URN: urn:nbn:se:lnu:diva-37167
73. Moja L, Kwag KH, Lytras T, Bertizzolo L, Brandt L, Pecoraro V, Rigon G, Vaona A, Ruggiero F, Mangia M, Iorio A, Kunnamo I, Bonovas S. Effectiveness of computerized decision support systems linked to electronic health records: a systematic review and meta-analysis. *Am J Public Health*. 2014;104(12):e12–22.
74. Garg AX, Adhikari NK, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *JAMA*. 2005;293(10):1223–38.
75. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ*. 2005;330(7494):765.
76. Black AD, Car J, Pagliari C, Anandan C, Cresswell K, Bokun T, McKinstry B, Procter R, Majeed A, Sheikh A. The impact of eHealth on the quality and safety of health care: a systematic overview. *PLoS Med*. 2011. Jan 18;8(1):e1000387. <https://doi.org/10.1371/journal.pmed.1000387>
77. Nachtigall I, Tafelski S, Deja M, Halle E, Grebe MC, Tamarkin A, Rothbart A, Unrig A, Meyer E, Musial-Bright L, Wernecke KD, Spies C. Long-term effect of computer-assisted decision support for antibiotic treatment in critically ill patients: a prospective 'before/after' cohort study. *BMJ Open*. 2014;4(12):e005370.
78. Berner ES, editor. Clinical decision support systems. New York: Springer; 2007.
79. Begum S, Ahmed MU, Funk P, Xiong N, Folke M. Case-based reasoning systems in the health sciences: a survey of recent trends and developments. *IEEE Trans Syst Man Cybern Part C Appl Rev*. 2011;41(4):421–34.
80. Khosla V. Technology will replace 80% of what doctors do. *CNN*, 4 December 2012.
81. Bergmann, et al. Case-based reasoning – introduction and recent developments. *Künstl Intell*. 2009;1:5–11.
82. Dehghani Soufi P, Samad-Soltani M, Shams Vahdati T, Rezaei-Hachesu S. Decision support system for triage management: a hybrid approach using rule-based

- reasoning and fuzzy logic. *Int J Med Inform.* 2018;114: 35–44.
83. “HL7 CDS Standards”. HL7 CDS Working Group, 2 August 2019.
84. Spie. Tanveer Syeda-Mahmood plenary talk: the role of machine learning in clinical decision support. SPIE Newsroom, March 2015.
85. Wagholarikar K, Sundararajan V, Deshpande A. Modeling paradigms for medical diagnostic decision support: a survey and future directions. *J Med Syst.* 2012;36(5):3029–49.
86. The thesis author’s own term, not widely used, but inserted here for pedagogic effect.
87. Mitchell M. An introduction to genetic algorithms. Cambridge, MA: MIT Press; 1996. ISBN 9780585030944
88. Balamurugan R, Natarajan AM, Premalatha K. Stellar-mass black hole optimization for biclustering microarray gene expression data. *Appl Artif Intell Int J.* 2015;29(4):353–81.
89. Bianchi L, Dorigo M, Gambardella LM, Gutjahr WJ. A survey on metaheuristics for stochastic combinatorial optimization. *Nat Comput.* 2009;8(2):239–87.
90. Boyd SP, Vandenberghe L. Convex optimization. Cambridge: Cambridge University Press; 2004. p. 129. ISBN 978-0-521-83378-3



Introduction to Artificial Intelligence in Medicine

5

Bart M. ter Haar Romeny

Contents

Introduction	76
The Development of the AI Framework	77
How Does a Convolutional Neural Network Work?	77
Convolution and Cross-Correlation	78
A Close Look into “AlexNet”	79
Representation Learning	81
Unsupervised Learning and a Geometric Model	81
Network Topologies, Types of Learning and Performance Measures	85
Topologies of Networks	85
Types of Learning	86
Measures of Performance: Sensitivity, Specificity, ROC	87
Inference and Network Examples	87
Deep Neural Network Application Domains	88
Relation to the Visual System	90
Color Processing and Colorization	90
Foveated Vision	91
Discussion	94
Lessons for All Doctors	94
References	94

Abstract

To better understand the mechanisms of the seemingly “black box” of AI and deep learning, we take a closer look at its internal processes. We will discuss the power of contextual processing, study insights from the human visual system, and study in some detail how different deep convolutional neural networks work. We do this with an engineering view, for radiologists, in an intuitive way.

B. M. ter Haar Romeny (✉)
Department of Biomedical Engineering, Eindhoven
University of Technology, Eindhoven, the Netherlands
e-mail: B.M.terHaarRomeny@tue.nl

Keywords

Deep learning · Convolutional neural network · Visual cortex · Visual learning · Context · Receptive fields · CNN · Brain efficiency · Visual pathways · Principal component analysis · Geometry · Recognition · Reconstruction

Introduction

The revolution in Artificial Intelligence in Medicine has surprised us, fascinated us, sometimes frightened us, and mostly benefitted us. Reports on spectacular performance abound, in many different application areas. Virtually all medical imaging and medical image analysis conferences (such as RSNA, ECR, and MICCAI) are now dominated by deep learning applications and papers, and new start-ups announce themselves every day.

This chapter gives a compact overview of the basic principles of AI in Medicine, specifically on deep neural nets, i.e., networks inspired on the adaptive synaptic connections in the brain. This chapter focuses in particular on an intuitive understanding of the technology and functionality inside such a network, with an emphasis on image understanding, to make the system less a magical trick. A section explaining the general pipeline and how the system actually learns is included, as well as a geometric model and how data are optimally represented. As modern neuroscience sees the same revolution, a section is dedicated to the relation with (and lessons from) findings in the human visual system.

The performance (anno 2020) of AI in different application areas is not the same:

- *Recognition* is handled very well: i.e., learning from large datasets to detect very subtle features, often outperforming humans
- *Reconstruction* to some extent: i.e., generating new data, such as CT data from MRI, generating new texts
- *Reasoning* only marginally: i.e., solving complex puzzles, making full differential diagnoses.

In general, much of the internal workings of deep neural nets is not known and they are still largely considered a “black box,” despite huge efforts to work on its explainability [1]. The jargon language of developing and coding deep learning networks is not easy to understand by other fields, such as medicine, biology, and neuroscience. Mechanisms for the distribution of (often open source) AI computer code such as the GitHub repository is very popular, as can be seen by their numbers: as of January 2020, GitHub has over 40 million users and more than 190 million repositories (including at least 28 million public repositories), making it the largest host of software source code in the world. It is promising to see how more and more AI specialists and medical doctors are working in teams together on many different dedicated AI applications.

This chapter is organized as follows: In section “[The Development of the AI Framework](#)” we describe history of AI and neural networks, and the winning configuration: *deep convolutional neural networks* (CNNs). The internal connectivity structure, training, and inference (actual use) are globally explained in section “[How Does a Convolutional Neural Network Work?](#).” section “[Representation Learning](#)” will discuss the mechanism of how learning can be modeled, with a bit of mathematics. Grasping some mathematics may make these systems less enigmatic. The different types of neural nets, methods of learning, and how the performance is measured is discussed in section “[Network Topologies, Types of Learning and Performance Measures](#).” In section “[Deep Neural Network Application Domains](#)” several application areas are discussed, with examples of the exploited network topology, and the current state-of-the-art in performance. The last section “[Relation to the Visual System](#)” focuses on human visual perception and its relation to these deep CNNs. Modern brain research has seen the same revolution, and may inspire not only a possible explainability but also lead to substantial future increases in speed and performance. The chapter is concluded with a discussion and a section with “Lessons for All Doctors.”

Every italic term can be looked up in Wikipedia for further study.

The Development of the AI Framework

Neural nets have been around for decades, but their performance was disappointing. One early type of AI were so-called expert systems, best explained with a famous example. In the 1980s the Campbell Soup factories in the USA exploited highly complicated sterilizers, for thousands of cans of soup. Not many people knew the intricacies, but one exceptional expert in troubleshooting named Aldo Cimino, who flew from plant to plant, retired. They decided to “can” all his knowledge in hundreds of rules. As this set of rules kept growing, it soon became unmanageable, and it failed. The first neural networks, also developed in those days, were limited to three layers (input layer, hidden layer, and output layer), and also disappointed in performance. It led to the so-called AI winter in the 1990s.

The last decade has seen the application of *pattern recognition* [2, 3], where large sets of hand-crafted filters (often up to hundreds) are designed to extract features from images. Typically sets of ground-truth images are processed in a pipeline, concatenating stages for preprocessing, segmentation, feature extraction, and feature selection. From these features then a high-dimensional *feature space* is constructed, in which apparent clusters are optimally separated by so-called *classifiers*. Famous classifiers are *support vector machines*, *random forest classifiers*, and *Bayesian classifiers*. There are dozens more. Despite the many sophisticated filter banks and clever classifiers, the performance was never as good as human experts, and one of the problems was that the hand-crafted design of features was either never complete or too redundant.

The breakthrough came with the realization that the features should be *learned* from the data. This turned out to be best done with neural nets with many more layers. The greater depth gives these network the name “deep.” The universal toy example for any AI student entering the field is the famous example of the neural network “LeNet.” This network, with only 7 layers, gave until then unprecedented performance in classifying images of handwritten digits. It was published by Yann LeCun et al. in a landmark paper [4]. The “MNIST” dataset of written letters (60,000

images of 28×28 pixels) is so small that the network can be handled and trained on any computer. In those days, the famous *ImageNet Large Scale Visual Recognition Challenge (ILSVRC)* was held annually, a competition to classify 1.4 million images into 1000 classes [5]. In 2012 Alex Krizhevsky et al. [6] (85000 citations) beat the competition with a staggering 12% better performance by introducing “AlexNet” with more convolutional layers (explained in section “[How Does a Convolutional Neural Network Work?](#)”), in essence an extended version of the “LeNet” network.

Several recent overviews discuss the history, applications, and future perspectives of AI in medicine well [7–11]. With today’s speed of distributing information, in general three levels of publishing new work can be seen: (a) the classical publications in peer reviewed journals and conferences. This has a relatively long turn-around time, from months to sometimes even 1.5 years. (b) Much faster are ArXiv (for computer science), BioRxiv (for the biosciences), and MedRxiv (for the health sciences), where papers are “published” (and the work claimed) during the review process. This has become popular, especially for the computer- and AI-sciences (ArXiv distributes close to 1.8 million articles). A third level is popularization and getting broad attention through Medium.com, an open publishing forum with many categories and many easy-to-read tutorial contributions on AI (see, e.g., towardsdatascience.com).

In medical image analysis, *grand challenges* have become increasingly popular, and data from hundreds of such competitions (with public datasets, publications in high ranking journals and open source software) are available (e.g., [12]). Recently, the quality control process of these influential studies have received renewed attention [13].

How Does a Convolutional Neural Network Work?

The simplest deep CNN is a neural feed-forward network, a chain with many layers. The number of layers can vary over a wide range depending on the application, e.g., from 7 layers (a “shallow”

network) up to 150 layers (a “very deep” network), or even more.

Figure 1 shows this fundamental CNN layered structure. To the left is the *input layer*, where data are fed into the network. The *output layer* is on the right. The input could be for example an input image, a video, a document, or a series of instrument measurements. The many blue lines are the *weights*, the “synaptic connections,” which are not known in the beginning, and are to be learned. In very deep neural nets, the number of weights can run into the hundreds of millions. The output could be, for example, a classification, or a translation, a modified image, etc. In section “[Network Topologies, Types of Learning and Performance Measures](#)” an overview is given of the many application areas where CNNs have been successful.

Many good introductory and overview texts are available [14–16], as well as many tutorials and downloadable implementations. In this section the main principles of a CNN are summarized, and a concise explanation is given of some

specific terminology that accompanies deep learning, for medical experts as primarily users of this technology.

Convolution and Cross-Correlation

Why is the network called “convolutional”? To understand the notion of *convolution*, it is instructive to study classical *template matching* in images. Suppose one needs to find the positions of the letters “a” and “e” in a piece of text given as an digital image (see Fig. 2).

The “template” is the small image that forms the letter “a” (e.g., 10×10 pixels), which is shifted row-by-row over the pixels of the much larger text image. Every time the template is on top of an “a” in the text, the *cross-correlation* will be maximal: *cross-correlation* is defined as the summed product of every pixel in the template and the underlying pixels of the text image. At other non-matching locations, the value of the correlation will be lower. *Convolution* is the same as

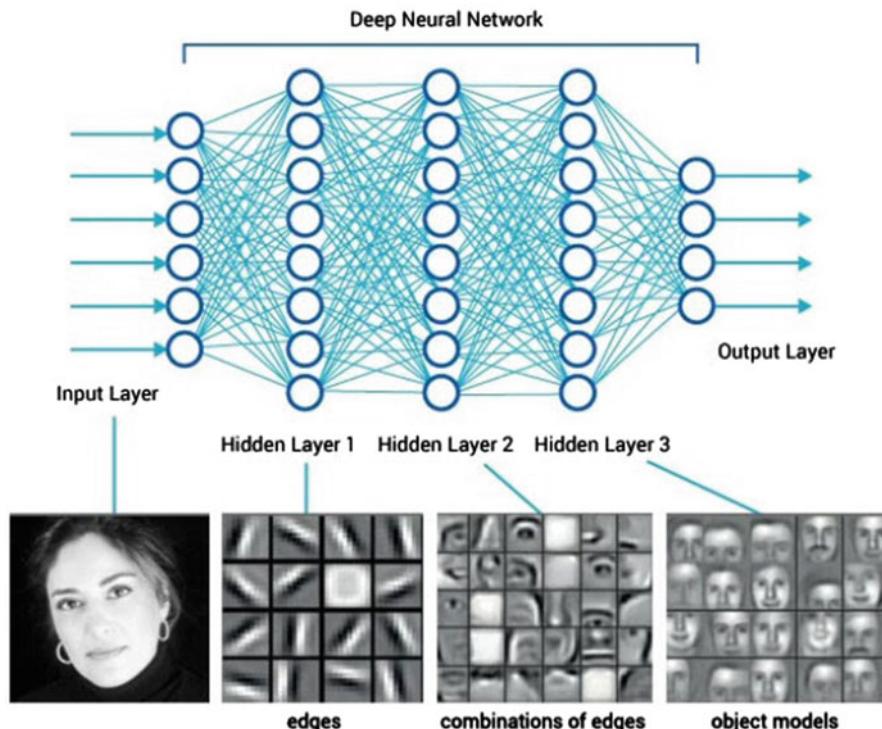


Fig. 1 A deep neural network

Fig. 2 Template matching
The templates (= filters = kernels) of an “a” and “e” (enlarged here for visibility) are shifted row by row over the discrete input image (top). At a match with the underlying image, a high cross-correlation is found. The template matches are shown in the output image as circles (bottom)



correlation, except that the template is then mirrored (in both the x- and the y-direction). Because convolution has specific mathematical advantages, convolution is the standard for neural nets.

The step size of the filter over the image is called the *stride*. Typically the stride = 1, but it can be larger, e.g., stepping every fourth pixel exploits a stride = 4.

Other names for “template” are “filter” or “kernel,” all denoting the same thing, so convolution is another name for “filtering.”

In the CNN the convolutional layers do convolutions. In the first layer the filters are very simple, and detect simple features (like the “a”), i.e., local edges and small line segments.

pixel. This image is convolved in the first convolutional layer with filters (kernels) of $11 \times 11 \times 3$ pixels. Typically the filters are square and have the same number of colors (in this case 3: RGB) as the image. Because the stride = 4, there are 55 steps of the filter over the image (in both the x- and y-direction), so the convolved output image of the first layer has a size of 55×55 pixels. A convolution is almost always followed by a threshold operation, whereby negative values are set to zero with a so-called *rectifying linear unit* filter (ReLU).

The second layer: because there are 48 different filters applied in the first layer, we get a stack of 48 convolved and thresholded images. This stack is a “multi-dimensional matrix” or *tensor* of size $55 \times 55 \times 48$ (actually the authors split the work over two graphical processing units (GPUs), therefore a second pipeline is partially seen above the first one in Fig. 3, so the total number of filters applied to the first layer is 96, but this is not relevant now). To prevent a true explosion of data, some form of data-reduction is done: *max pooling*: only the datapoints with the largest values are kept over a small shifting area, in this case by taking the maximum value in a shifting $3 \times 3 \times 48$ cube with stride = 2. Often this is done explicitly in a separate layer, the *pooling layer*, but here integrated in the second layer.

This output of the pooling layer is then fed into the next convolutional layer, where the data is convolved with 128 kernels of size 5×5 with stride = 2, so the size of the output tensor is

Forward Pass

The input image to the left is a color image with a size of 224×224 pixels with 3 colors (RGB) per

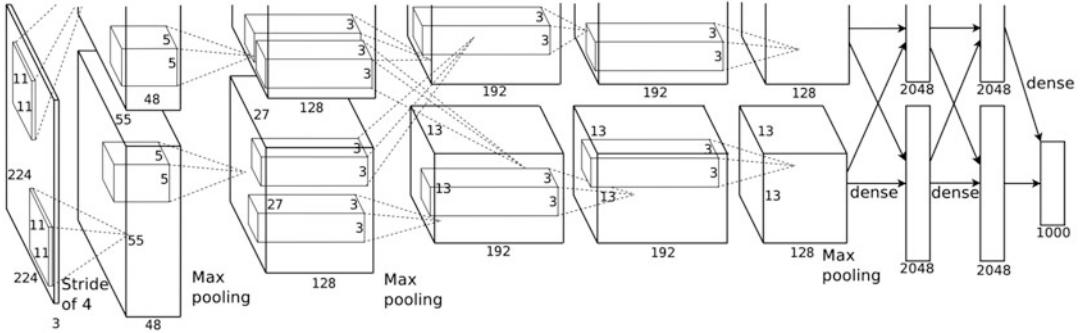


Fig. 3 AlexNet. (Figure from [6])

$27 \times 27 \times 128$. Max pooling again, and convolution with 192 filters of size 3×3 , etc. The images get progressively smaller, and contain less detail but more global information.

The final stage contains the so-called *fully connected layers*. They function as a *classifier*, in this example network to get 1000 neurons (classes) as output. They do this by matrix multiplication with a rectangular matrix, to change the dimensionality. In this way the final tensor of $13 \times 13 \times 128$ is converted into a vector of 1000×1 , which gives the probabilities of the 1000 output classes.

Two more layers are needed: a dedicated layer at the input and a dedicated layer at the output. Because a neuron in the computer can be only be fed with numbers, the input has to be converted to numbers when they are not-numbers, e.g., characters or classes. Such a converter is called an *encoder*. An encoder does any preprocessing, and may include resizing of input images, proper alignment, color mapping, etc. For the output the same holds: output numbers have to be converted to, e.g., classes (“17” = class “cat”). The converter at the end is called the *decoder*. When the output has to be interpreted as a series of probabilities which together sum up to 1, a special *softmax function* is applied in the decoder.

Backward Pass

But how does a neural network learn? After the forward pass follows the backward pass. Learning happens in the backward pass with a process called *error back-propagation*. Initially, all the weights in the network are unknown, and

initialized as random numbers between 0 and 1. They are the blue lines in Fig. 1. When an input image is processed in the forward pass of this random network, it gives a random output, which is likely to be wrong. The weights are then adjusted from the last layer onwards to the first, until the error (or *loss function*) is minimum. So when, for example, a digit recognizing CNN outputs a “5” when the input was a picture of a hand-written 6, the error is +1. When the output is a set of probabilities, an often used *loss function* to estimate the error is the *cross entropy function*. Then the process repeats itself: the next training image is presented and processed in the forward pass, and the weights adjusted in the backward pass. And so on with enormous amounts of training images, until the overall error is sufficiently small. This is the *training process*.

This adjustment by backpropagation is a lot of work for the often millions of weights. The process is well understood. Mathematical optimization techniques like *gradient descent* (to find the minimum, always head downwards) and the use of special hardware such as graphical processing units (GPUs, game cards) have made this finally feasible. During the training, the performance of the net is constantly monitored by frequent tests (e.g., every 100 or every 1000 passes) by doing forward passes with dedicated *validation data with ground truth*. Of course these validation data should be different from the training data.

When the training is finished, the final performance of the system is established as an unbiased evaluation with a *test set with ground truth*.

Often it is difficult to get enough training data. It turns out that small variations of the input image (small rotations, warpings, zoomings, shifts, contrast changes, etc.) can act as “new” unseen input images. This process is called *data augmentation*.

The shaping process of the weights during the training can be visualized. Andrej Karpathy, then PhD student at Stanford University and currently Sr. Director of AI at Tesla, built an instructive website where the CNN can be run in a browser [17]. The convolved outputs of the respective network layers and the momentary error rates of the network can be seen evolving during the training process, for a number of simple networks.

Representation Learning

A crucial question that any user of AI has is: “How does a machine learning or deep learning model make its decisions?”. Or: “What happens inside the black box?”. This is a field of extensive research today, with still partial answers, mostly heuristical. The high complexity and the enormous number of connections do not make this an easy endeavor. By far the greatest percentage of published papers and programs on AI are applied, while leveraging theoretical foundations is still in the beginning. See for a good state-of-the-art overview the paper by Xie et al. [1], which discusses the *explainable AI* problem from many viewpoints.

The similarity to mechanisms of human visual perception is still marginally treated. The fields of modern AI computer science and modern neuroscience research seem still largely separated. The number of papers in both fields has exploded, with relatively few papers focusing on the cross-fertilization. Possible reasons are the broad difference in language (programming/modeling/mathematics versus physiology/pharmacology/genetics), the lack of computational tools for many vision scientists, and decades of historical modest interactivity.

One approach that combines the mathematical and human vision point of view is geometry. We treat one example in the section below. The treatment is some-what technical, but it pays to

understand some of the physics and mathematics of at least the first stages of a CNN (and human vision).

Unsupervised Learning and a Geometric Model

When we want to begin to understand why the self-organization works, we need to know how complex data, like large datasets of medical images, are optimally represented. Optimal in the sense that they take the least space (handy for storage, transfer, and computing) and still retain all information.

In order to begin to understand the stepwise increments in contextual processing exhibited in both deep neural networks and the visual cascade (remember the stages in Fig. 3), it is good to start in the smallest possible environment: a single pixel. Our notion of a pixel is best visualized when we zoom in, and start seeing the pixels. Invariably, these are squares, see Fig. 4 (left) of the author’s face.

But this pixel shape introduces “spurious resolution” [18], i.e., false edges and corners that are certainly not there. Square pixels are easy to fabricate, but the square shape is intrinsically wrong. Nature nowhere exploits squares. It can be proven, from an elegant reasoning with first principles, that the optimal shape of a pixel is the Gaussian kernel (a ubiquitous kernel, the same shape as the *normal distribution*). A Gaussian pixel shape has smooth boundaries (no sharp edges) and is circular (no sharp corners).

So the world is best observed with Gaussian-shaped pixels. In mathematical words: such an observation, where pixels slide over the image, is a convolution, as was discussed in section “Convolution and Cross-Correlation.” Mathematically this is written as

$$obs = pG \otimes w \quad (1)$$

where obs is the observation, pG is the pixel with a Gaussian shape, w is the outside world that is looked at, and the symbol \otimes denotes the convolution operator.

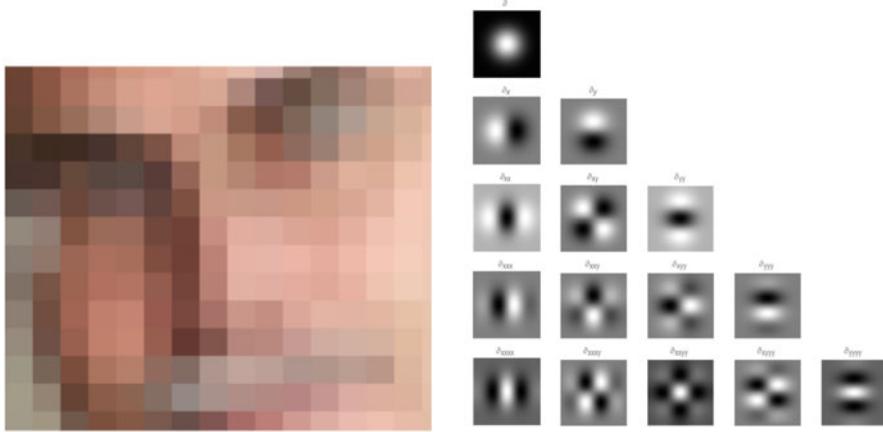


Fig. 4 Left: Square pixels in the zoomed face of the author, generating “spurious resolution,” i.e., false edges and corners. Right: The set of Gaussian derivative filters till 4th order

The next context level is to look at the direct neighbors of the pixel from the previous paragraph. This is done by looking at the difference with the pixel, mathematically described as the *derivative*, indicated with $\frac{\partial}{\partial x}$. The first derivative is the difference in intensity with the next-door pixel, and denotes the local edge strength.

Interestingly, when we take the derivative of our observation (to, e.g., detect edges), we get

$$\frac{\partial}{\partial x} \text{obs} = \frac{\partial}{\partial x} (pG \otimes w) = \frac{\partial pG}{\partial x} \otimes w \quad (2)$$

where $\frac{\partial pG}{\partial x}$ is the derivative of the Gaussian pixel. So edges are measured by convolving (filtering) with the derivative of a Gaussian. This can be expanded to the second derivative (i.e., the derivative of the derivative), third and higher derivatives. The set of all Gaussian derivatives in the x - and y -direction to fourth order is depicted in Fig. 4. On top is the Gaussian kernel (actually the “zero-th derivative”), on the first row the filters for edge detection in the x -direction (i.e., for vertical edges), resp. in the y -direction (for horizontal edges). The higher order derivatives are more complex filters, used for more complex measurements. For example, the second order derivative is involved in measuring the curvature of contours.

These filters are excellent models for the filters that are measured in the first layer of a fully trained neural net [6], and in the first layer of the

visual cortex [19], see Fig. 5. They measure the local *geometry*, i.e., the immediate contextual neighborhood around each pixel. It is interesting that only low order derivatives (up to maximally 5th order) appear in the first layer of both neural networks as in the visual cortex.

There is a mathematical paradigm, called the *Taylor expansion*, that says that we can describe a neighborhood around a pixel with excellent approximation as a (truncated) series of weighted higher order derivatives. See for a clear and entertaining visual explanation of the Taylor expansion [20] (video), and Fig. 6.

In literature these filters are often modeled with so-called *Gabor kernels*, but these, though very similar in shape, do not describe the local *differential geometry*.

So far the mathematical model. We can also do actual measurements of many of such local contextual neighborhoods around a pixel, and find their best representation. In Fig. 7 we have taken numerous patches (in this case of just 12×12 pixels) from a CT slice image. A famous mathematical technique to find the optimal representation is *principal component analysis* (PCA), a procedure that the computer easily can perform on these hundreds to thousands of patches. Figure 7 (top row) shows the PCA result for patches taken from a CT slice: the first 25 basis-patches (these are the so-called *eigenvectors* or *eigenpatches*). Note that these eigenpatches excellently agree with the

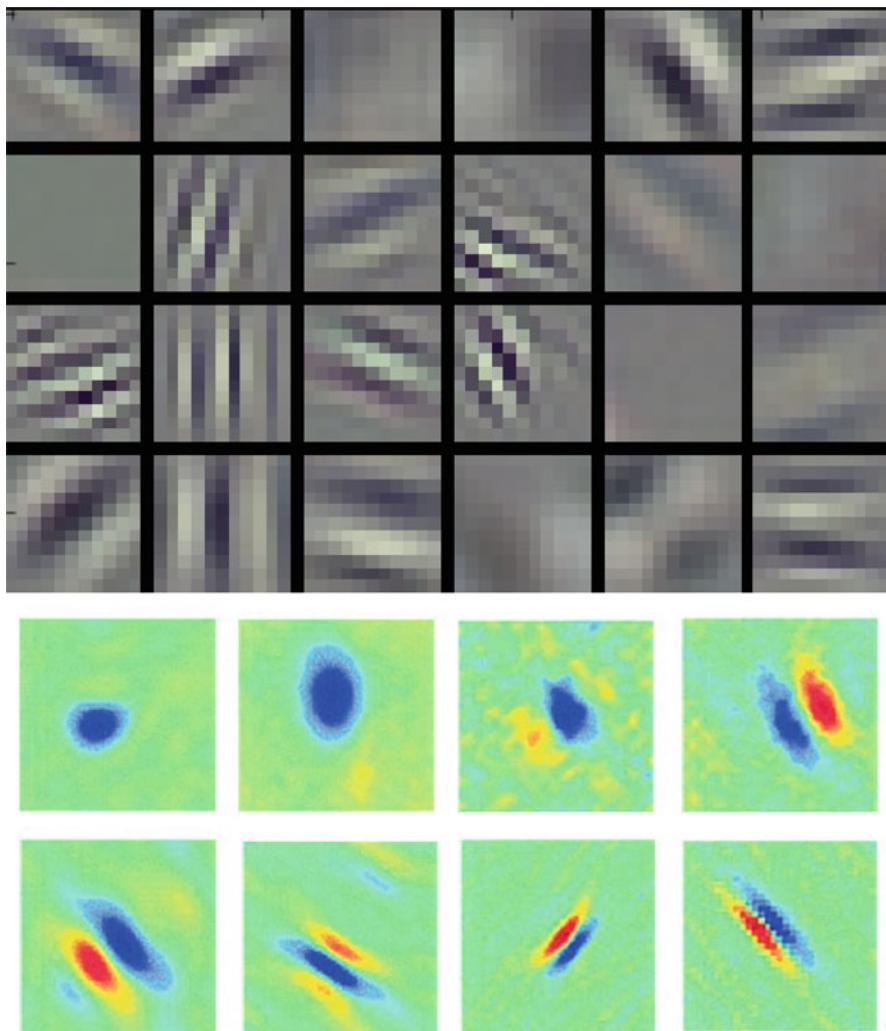


Fig. 5 Top: Learned filters in the first convolutional layer in a fully trained AlexNet. (Adapted from [6]). Bottom: Filters (receptive fields) of so-called simple cells measured in macaque primary visual cortex V1. (Adapted from [19])

model of Gaussian derivative filters described above.

From these eigenpatches *any* patch can be constructed (see figure 6). This is highly efficient: any given patch can be represented as a weighted sum of a small set (20 or less) of these eigenpatches, and we only need to store these basis patches. See how closely these measured data resemble the Gaussian derivatives of the previous paragraph. It is likely that the first layer in a CNN and the first visual layer in our visual system do just that. The eigenpatches form a so-called *basis*, and any patch can be represented with this basis.

This is why deep learning is often called *representation learning*.

With learning different data, different filters emerge. So patches from an image with restrictions, e.g., a picture with mainly vertical trees, leads to different eigenpatches, see Fig. 7 (bottom row). Now the first eigenpatches only measure vertical edges, as there are virtually no horizontal edges in the input image.

An example/proof comes from biology: in 1976 a famous experiment by Blakemore et al. [23] (with video) showed that a kitten, trained from birth with only seeing horizontal bars during

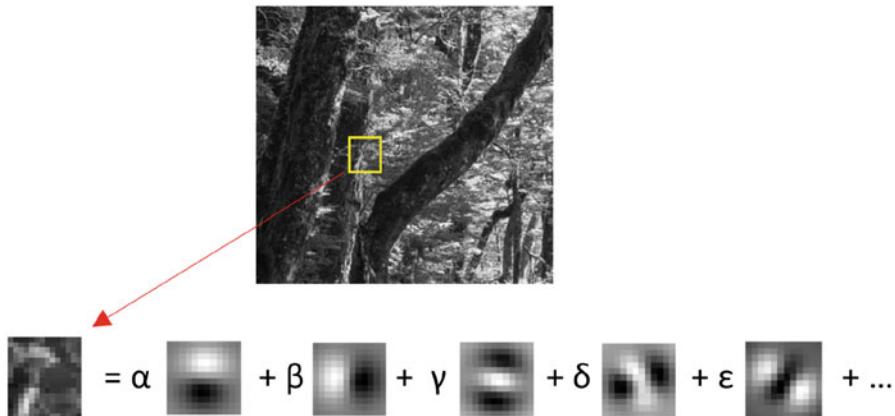


Fig. 6 Every signal, also 2D images, can be described as a sum of basis filters, and approximated with a finite number of basis filters, often < 20 . The small image patch is well

described by a series of image derivative filters. The Greek letters are the weights

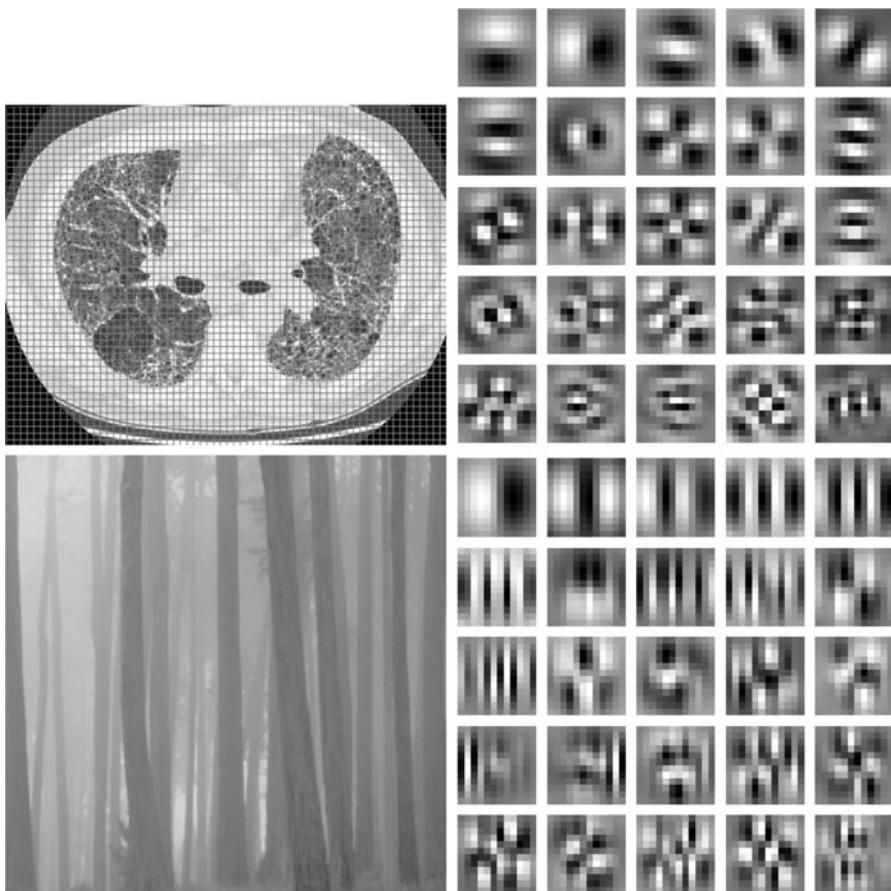


Fig. 7 Top row: Patches are sampled from a CT slice. Right: the first 25 eigenpatches. Bottom row: when patches are taken from an image with restrictions (e.g. only vertical trees), the resulting first 25 eigenpatches (right) show only

horizontal derivatives to high order for the first 7 patches, as vertical derivatives (for horizontal edges/contours) are hardly present in this image. (Adapted from [21, 22])

the first 3 months of its life, could completely not see any vertical bars after this period: it had not developed vertical edge detectors in its visual cortex, as vertical information was not present in the images it saw. See Fig. 8.

It is interesting that our vision also shows this phenomenon: the faces visible in Fig. 9 cannot (or with much more effort) be recognized when the same image is presented upside-down.

It turns out that most natural scenes have a wide variety of edges in many orientations, so this broad set of learned filters in the first stage is often the same for many application areas. It explains why *transfer learning* often works: start with a trained deep neural net from one domain (e.g. skin cancer detection) and retrain it on images from another domain (e.g. trained on ILSRC data). This turns out to substantially reduce the training time, as the filters in the first

layers do not start from a random initialization, but are already close to the required filter shape.

Network Topologies, Types of Learning and Performance Measures

Topologies of Networks

The number of network topologies (layouts) has greatly expanded, and is still growing. It is beyond the scope of this chapter to discuss these network topologies. See, e.g. [25] for an overview of CNNs in radiology. The “Neural Network Zoo” of the Asomov Institute [26] exhibits the many frameworks, see Fig. 10. Well-known networks are: LeNet (1998), AlexNet (2012) (discussed in more detail below), VGGNet (2013), Inception v3 (2015), ResNet (2015), DenseNet (2016).



Fig. 8 Blakemore’s cat. From birth, this kitten was shown – only – horizontal lines during a period of 3 months. After that period it could not perceive a

horizontal stick but it could well see a vertical stick. This kitten had not developed (i.e., learned) any receptive fields for vertical edges. (From [23] (see also the video))

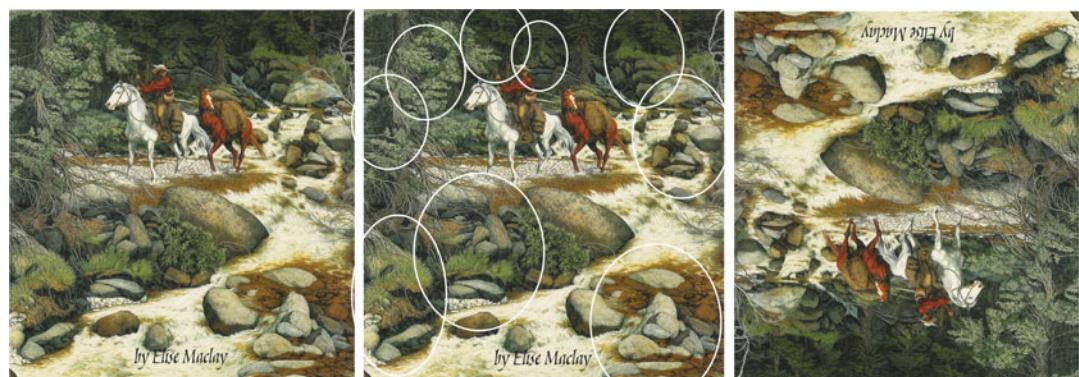
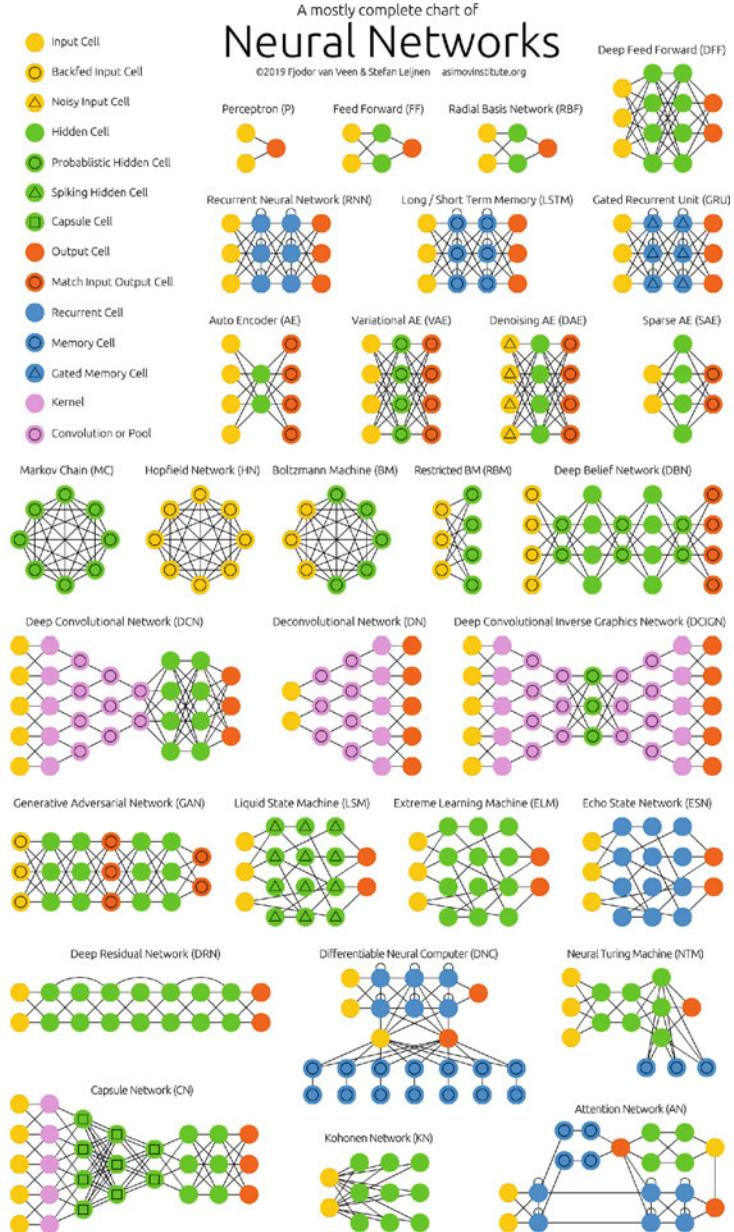


Fig. 9 Face recognition: the faces in the forest (locations indicated by circles in the middle picture) cannot or hardly be recognized in the same upside-down picture to the right.

We have never learned faces that are presented upside-down. (From [24])

Fig. 10 The “Neural Network Zoo,” a mostly complete chart of neural networks. (Figure from [26])



Types of Learning

In grand view, there are three main types of learning:

- *Reinforcement learning*: An agent tries to maximize the total amount of reward it receives while adaptively navigating through a set of predefined behaviors and rules until a goal is achieved [27].

- *Supervised learning*: The training is done with many ground truth examples, which often involves extensive annotation labor. The majority of the networks today are trained with supervised learning.
- *Unsupervised learning*: Where no teacher is necessary, all is trained from the raw data. This process, so natural for humans, is still largely enigmatic and is focus of intense research.

Measures of Performance: Sensitivity, Specificity, ROC

To trust the decisions made by AI systems, a quantitative measure of performance is necessary. These are specified in the conventional way for validation of a decision with specified measures for *sensitivity* and *specificity*, a *receiver operating characteristic* (ROC; see for a tutorial for computer-aided diagnosis systems [69]), a specification of the exact data the network was trained on, and the volume of the training data and the test data with ground truth.

Of course, networks are not perfect. The training set may not be sufficiently large, the algorithm reproduces what it has learned, the chosen topology is not optimal, etc. Sometime inadvertently errors occur. A tutorial example was presented by Narla et al. [28] (as a correction on their highly cited Nature paper [29]) that algorithms are more likely to interpret skin lesion images containing rulers as malignant because images with rulers were more likely to be malignant in the training data, see Fig. 11. Another example is the estimation of the age of a person, given an image of his/her face. A network trained on the public IMDB-WIKI dataset (with over half a million faces [30]) gives an erroneous result when the face is presented with too much surroundings, see Fig. 12, for which it was not trained.

The best insight in deep learning performance gives the direct comparison against healthcare professionals, in the specific topical domain. See for an extensive literature review [32].

Inference and Network Examples

Training is the expensive phase, and might take days, or even weeks to accomplish for very deep networks and millions of input data. Once done, a fully trained network is as an expert brain, the “golden egg,” and works fast (this could even be in milli-seconds). Application of a trained network is called *inference*.

For those not programming deep neural networks: many trained classical deep CNNs can be downloaded. They can subsequently be used for their original purpose, used for transfer learning with input from other domains, or to inspect what happens inside. It is common to see papers appear with full instructions on from which CNN to start, how to adapt and with downloadable free training data (e.g., [33]).

There are numerous free collections of downloadable working deep neural networks, like the TensorFlow Model Garden [34] and the Caffe Model Zoo [35]. For nontechnical users, a highly tutorial collection (directly executable with the software environment

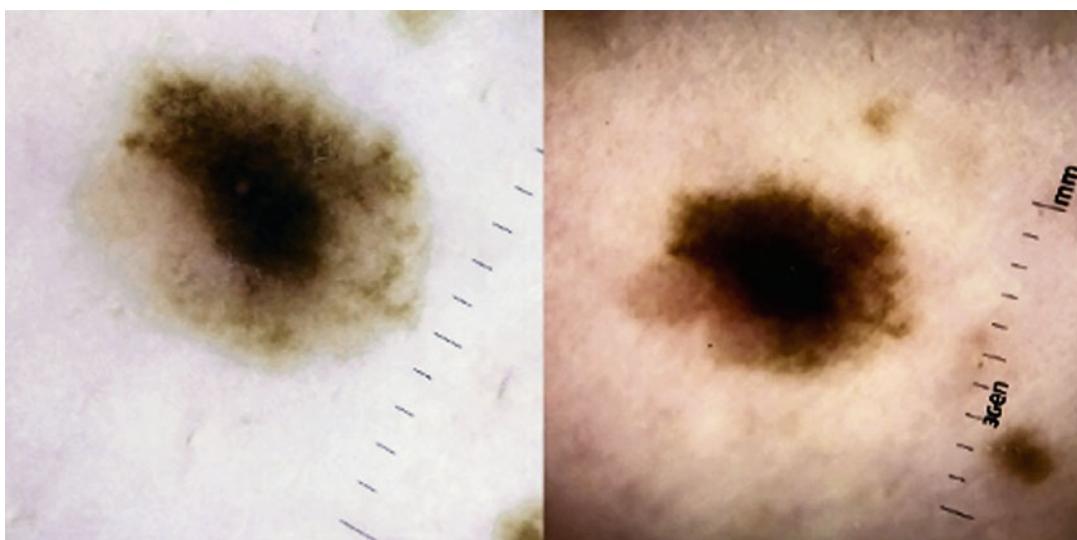


Fig. 11 Skin lesions with rulers can lead to an inadvertent bias towards malignancy. (Figure from [28])

Fig. 12 Age detection from a face [30] fails when the surrounding context is much larger than just the cropped faces of the training set. (Figure from [31])



Mathematica) is the Wolfram Neural Network Repository [36].

Deep Neural Network Application Domains

The impact of AI and deep learning grows rapidly and successfully. Below a concise and in nature incomplete set of application domains is presented, with some example references (also non-medical in order to give a broader overview of current directions) where deep learning is successfully applied. Many more applications are discussed in much more detail in the other chapters of this book.

- **Speech Recognition and Translation**

Well-known examples are Amazon Echo, Apple Siri, Google Translate, etc. These applications struggled for performance for decades, but now deep learning and their continuous training due to large-scale use (think here of millions of users) make them top performing today. The development of *natural language processing* in radiology has been lagging the imaging applications, but is now successfully taking off [37]. Note that recent language networks like GPT-3 not only have learned to autocomplete sentences pretty well (having learned from huge amounts of texts), but can do this even for computer code like *html* (the

language for browsers in which webpages are programmed).

- **Segmentation**

The prospect of self-driving cars became reality with the advent of real-time segmentation of traffic scenes [38] (with an illustrative movie). Many areas in medical imaging have seen successful segmentation [39], especially due to the invention of the so-called U-nets [40] (see Fig. 13), e.g., in the fields of brain MRI, breast MRI and cardiac CTA [41].

- **Detection**

AI has revolutionized histology, a classical area where computer-aided diagnosis was difficult [42]. Another example of successful automated detection is CAD4TB, a system for tuberculosis detection from X-ray images in situations of local shortage of doctors [43], see Fig. 14.

- **Classification**

A classic example (taken from a computer vision challenge) is age estimation from a single face image. Such a system can detect very subtle facial features that human experts cannot even name, when trained from 520.000 face pictures [30]. It inspired a deep learning model that could exploit the gap between predicted brain age using MRI and chronological age as a biomarker for early-stage neurodegeneration [44].

- **Image reconstruction**

Excellent results have been obtained in reconstructing fast dynamic sequences of 2D

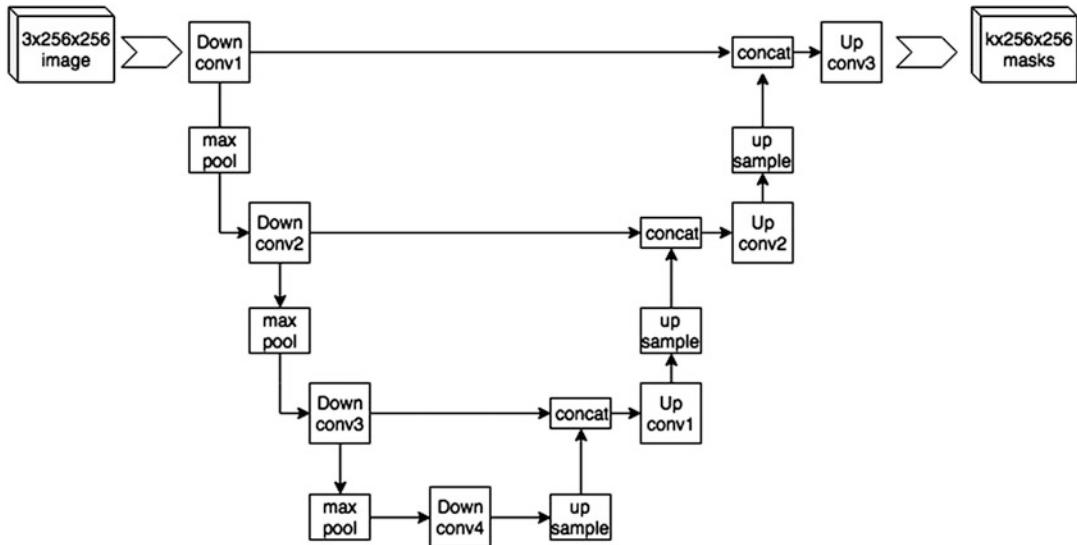


Fig. 13 A U-net [40]. In the left branch the chain of convolutions make the image smaller by a coarser representation (“down-sampling”), in the right branch, with

many side-connections, the resolution is increased again (“up-sampling”), leading to proper segmentation. (Figure from WikiMedia)



Fig. 14 CAD4TB: A lab operator in Ghana studies an AI-detected tuberculosis heatmap of a lung X-ray, acquired in a mobile van. Green and yellow areas point to a risk for abnormalities due to tuberculosis [43]

cardiac MR images from undersampled data using a deep cascade of CNNs [45].

- ***Super-resolution***

Deep learning methods on super-resolution MRI have been shown to generate 7T quality images from large sets of 3T MRI data [46] and was successful in super-resolution musculoskeletal MRI [47].

- ***Content-based image search and retrieval***

Examples abound in daily use such as Google Lens, Google Photos, Facebook, face tagging, etc. These developments are based on the huge numbers of images available (Google Photos: 1 billion users, 1.2 billion pictures are uploaded daily). It is an exciting future perspective to do radiological *query-by-visual-*

search in the ubiquitous PACS archives, e.g., the most similar images as the ones under study [48].

- ***Image restoration***

AI offers good performance in many image restoration tasks, such as *inpainting* (filling-in gaps in images with the contextual texture of the surroundings) [49], denoising, enhancement, and CT/MRI image quality improvement [50].

- ***Image registration***

Image registration (matching) is an ill-posed problem, and faces severe difficulties with conventional methods. Both supervised and unsupervised methods for direct and fast transformation prediction are hot topics in research, and the number of successful application areas in medical imaging is steadily increasing. See [51] for a recent review.

- ***Image, text and speech generation with GANs***

Generative adversarial networks (GANs) generate new images (or text/speech/video), by a competition of two neural networks. One network generates fake data, and the other tries to establish if it is fake or real. The results in generating new highly realistic images such as faces (e.g., “StyleGAN” face generation in a browser [52] (search YouTube for “StyleGAN2”)), but also “new” medical images used for data augmentation) are outstanding. GANs are also exploited to synthesize CT data from MRI data [53], of much practical use in radiotherapy. The growth of GAN applications is turbulent. See for a review on GANs in medical imaging [54].

- ***Image style transfer***

The advent of so-called cycleGANs, generative adversarial neural networks that propose a structure preserving loss in unpaired image-to-image translation frameworks [55], lets the network learn a specific style (e.g. by a painter as van Gogh) and transforms the input image into that style. StyleGANs have proven to be successful for staining normalization in digital histology images [56].

- ***Neural radiance fields***

Recently it has become possible, with a technique called neural radiance fields, to reconstruct complex outdoor scenes like

buildings in 3D with high precision from huge amounts of unstructured (“in the wild”) photographs. See: [57] and the impressive video.

Relation to the Visual System

Not many papers on AI make the comparison to human visual perception. The similarity of a deep CNN with the visual pathway is beneficial in dealing with their explainability [58].

Vision is by far the most extensively studied brain function and many excellent state-of-the-art overviews exist [59–61]. The classic popular overview by Nobel Prize winner David Hubel [62] is still a good entry in the field. It turns out that the visual processing in our brain is massive: about one quarter of our brain’s 10^{10} neurons is involved in visual processing, we really are “visual machines.” The internet exploded when it was possible to handle images easily.

In vision we also recognize a cascade of layers in the occipital part of the brain, which are termed V1 (primary visual cortex), V2, V3, etc., see Fig. 15. A striking difference between the human and artificial networks is the difference in use of energy. A typical human brain is estimated to use about 25 Watt, while modern computing data centers use even Megawatts. Another difference is the frequency: neurons fire at rates not higher than 4–6 kHz, while current laptops have processors running at 3 GigaHz. The difference is so many orders of magnitude that the conclusion must be that still an enormous number of clever strategies can be learned from visual circuits.

A few examples as food for thought for energy saving are given below.

Color Processing and Colorization

In computer vision applications a color pixel is processed as a triple of red, green and blue (sometimes 4: transparency). This triples the computational costs. It is known that color processing in the visual cortex is more efficient: color-sensitive areas in the visual cortex are organized in *blobs*,

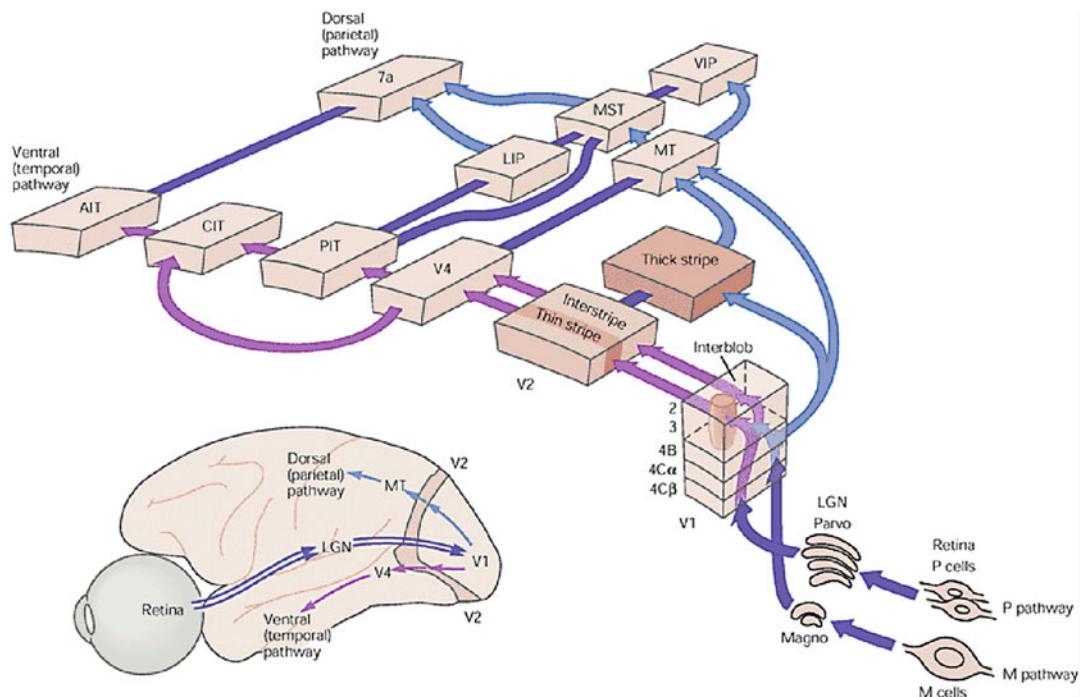


Fig. 15 The layers in the visual system. From the retina (lower right) two streams (parvocellular for shape and magnocellular for motion) project to the thalamus, from

where they project to the primary visual cortex V2, V2, V3, etc. From [62]

stainable with cytochrome oxidase, and take less than 10% of the available space. Maybe our conventional triple pixel processing is overkill? Levin et al. demonstrated that a grayscale image can be colored per segment by only giving some color hints [63], a process called *colorization*. See Fig. 16. AI neural nets can be specifically trained for color, such as ColorNet [64]. For the layout of the network see Fig. 17. This network fills the coloring into a back-and-white input image quite reliably. For an example see Fig. 18. In visual neuroscience it was found that such a mechanism, i.e., scene segmentation and the binding of color to object surfaces, seems to converge in V2 in the visual cortex [65].

Foveated Vision

Our retina is not just a simple camera. Having a fovea saves energy, as now only the 3D world is sharply painted in our internal representation

where we actually look, where we focus our attention. The rest is at lower resolution. First described in [66]. The illusion in Fig. 19 makes the point. Of the 12 black points present, only one can be seen at a time. Foveation was applied in a CNN by Ghafoorian et al. [67], using filters with a spatially changing resolution, sharp in the middle and blurred further out. See Fig. 20. As they reported, the performance was excellent. The kernel permitted a larger contextual area to be covered for less computational effort. This principle, of only painting the attentional focus into the internal representation, is also used in 3D *Simultaneous Localization And Mapping (SLAM)*. Here the fovea is replaced by a *Lidar*, an infrared laser range measuring system, to map the surroundings. This technique is ubiquitous in AI systems for self-driving cars. See [68] for an instructive movie.

Many more of such inspirations for energy savings are likely to be surfacing in the near future.



Fig. 16 Colorization. Given a grayscale image marked with some color scribbles by the user (left), the colorization algorithm using optimization produces a colorized image

(middle). For reference, the original color image is shown on the right. (Figures and caption text adapted from [63])

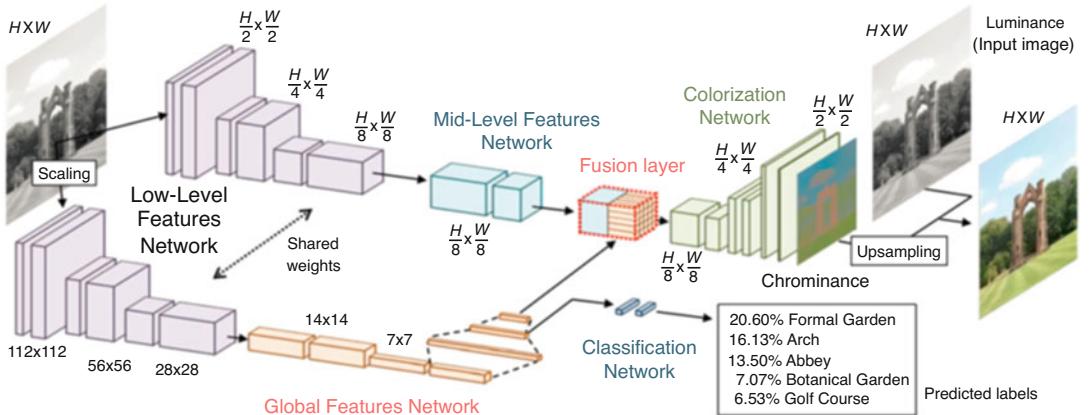


Fig. 17 Network layout of ColorNet. The initial layers deal with low-, middle-, and global features to come to a segmentation, and a upsampling stage is trained for filling in the color. (From [64])



Fig. 18 ColorNet [64]. Left: original. Middle: greyscale. Right: colorized by the neural network. Example taken from the Wolfram Neural Net Repository [36]

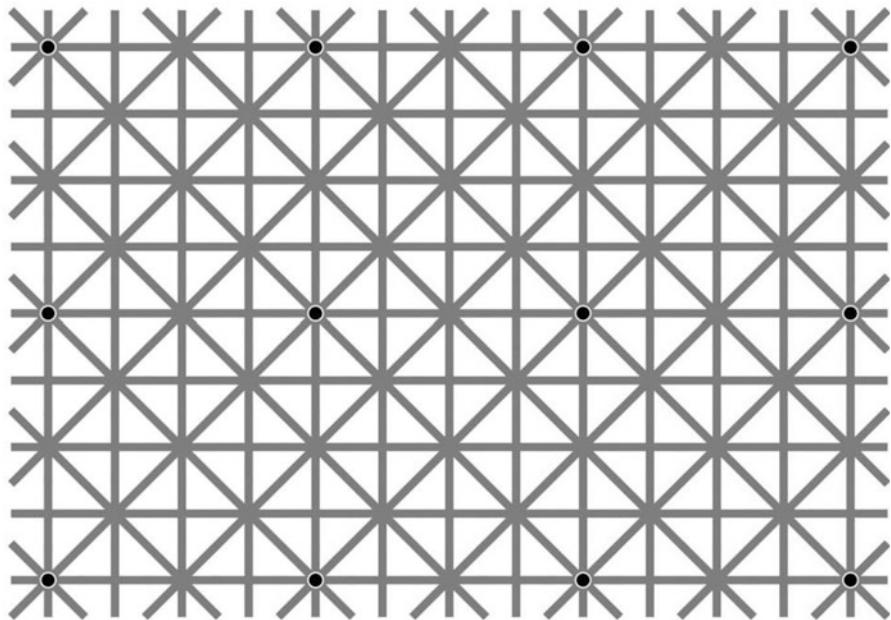
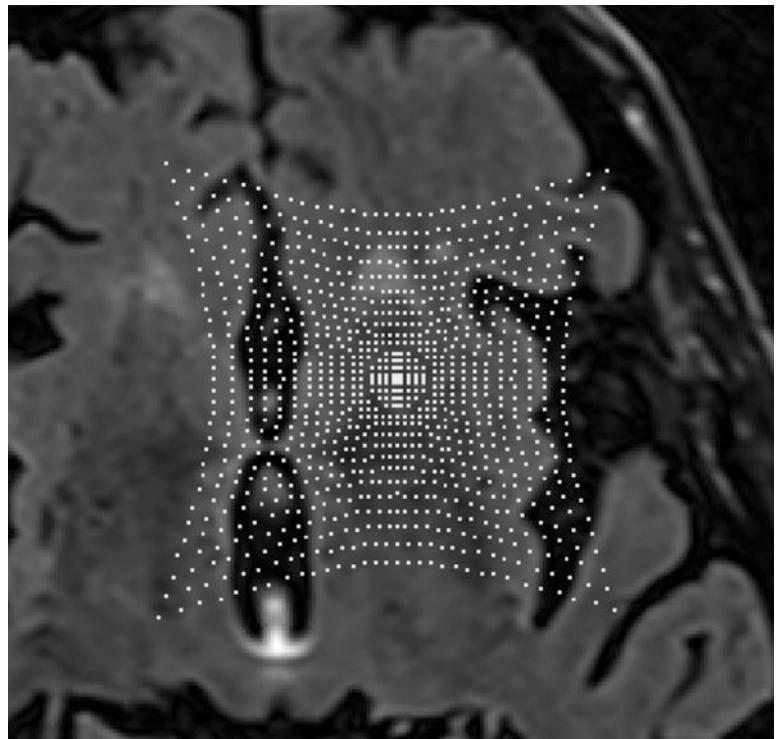


Fig. 19 12-dot illusion. There are 12 black dots, but only one can be seen, due to our foveated retina. (From [66])

Fig. 20 Foveated kernel applied as filter in a deep neural net. (From [67])



Discussion

In this chapter we looked more closely inside the basic structure and mechanisms of deep convolutional neural networks, the workhorses of modern AI for images. The purpose is to take the magic off such networks, to appreciate the methods they employ for learning and what is necessary for optimal performance. The developments go fast, and many new network types are being developed and exploited, so the treatment is only elementary. Especially in the fields of generative adversarial networks with its many variations (conditional, cycle-consistent, style-), unsupervised learning, ever increasing hardware performance, and the availability of big data substantial (often even spectacular) steps are taken.

Modern GPU technology has taken the place of regular computing in AI, i.e., massively parallel supercomputing performance is now available at competitive prices. For example, the Nvidia DGX-A100 GPU delivers 5 petaflops (5×10^{15}) or five quadrillion (thousand trillion) floating point operations per second), faster than the world's fastest supercomputer was in 2012, for 200K\$. GPU computing power can now be rented for reasonable prices in the cloud.

Currently much of the AI development is still energized by research groups, which involves open source software and freely available trained networks and datasets for evaluation. With the advent of networks that are just too big to install on regular computers (such as the language model *GPT-3*), and the advent of many more commercial companies (from start-ups to unicorns) active in the medical arena of opportunities, this is likely to change soon.

Lessons for All Doctors

- AI in Medicine is here to stay, as a major supportive medical device. The successes are already impressive in many application areas, and a rapid expansion is seen. As in any field, it remains imperative to be critical. A thorough understanding of AI's tools and their possibilities enables the acceptance, and the way how to control its introduction.

Collaboration between medical and computer scientists is key. In many pioneering hospitals, AI specialists are already embedded in the clinical divisions, often in large groups. Radiology and pathology are rapidly adopting the new AI tools. Such tools need to be trained specifically for the focus of that department, with own (often longitudinal) data. Data security also requires that the development is best done in close collaboration on-site.

- AI should be embedded in the training of medical specialists from the beginning. Today's tools are so user friendly, that even doctors can play with the principles on their laptop, train their own teaching networks, to grasp the look-and-feel of what these new tools do, and how a dedicated local database should be designed.

New paradigms are needed in AI research. As the brain is extremely energy efficient and the current implementations of AI systems for processing and storage are not, much inspiration is expected to come from biology. The worlds of AI vision and biological vision is fusing, and the cross-fertilization can have tremendous benefits. In that sense: AI has actually only just started.

References

- Xie N, Ras G, van Gerven M, Doran D. Explainable deep learning: a field guide for the uninitiated. arXiv preprint arXiv:2004.14545. 2020.
- Bishop CM. Pattern recognition and machine learning. New York, NY: Springer; 2006.
- Duda RO, Hart PE, Stork DG. Pattern classification. 2nd ed. New York: Wiley; 2001.
- LeCun Y, Bottou L, Bengio Y, Haffner P, et al. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86(11):2278–324.
- ImageNet. Large Scale Visual Recognition Challenge (ILSVRC), 2010–2017. ILSVRC evaluates algorithms for object detection and image classification at large scale: 150000 photographs, 1000 classes. <http://www.image-net.org/challenges/LSVRC/>
- Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. Advances in neural information processing systems 25. Red Hook, NY: Curran Associates; 2012. p. 1097–105.

7. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer.* 2018;18(8):500–10. <https://doi.org/10.1038/s41568-018-0016-5>.
8. Kaul V, Enslin S, Gross SA. The history of artificial intelligence in medicine. *Gastrointest Endosc.* 2020. <https://doi.org/10.1016/j.gie.2020.06.040>.
9. Mintz Y, Brodie R. Introduction to artificial intelligence in medicine. *Minim Invasive Ther Allied Technol.* 2019;28(2):73–81. <https://doi.org/10.1080/13645706.2019.1575882>.
10. Ranschaert ER, Morozov S, Algra PR. Artificial intelligence in medical imaging: opportunities, applications and risks. Springer; 2019. <https://doi.org/10.1007/978-3-319-94878-2>.
11. European Society of Radiology (ESR and others). What the radiologist should know about artificial intelligence – an ESR white paper. *Insights Imaging.* 2019;10(1):44. <https://doi.org/10.1186/s13244-019-0738-2>.
12. DIAG Nijmegen MEVIS Fraunhofer. Grand challenge, a platform for end-to-end development of machine learning solutions in biomedical imaging. 2020. <https://grand-challenge.org/>
13. Maier-Hein L, Eisenmann M, Reinke A, Onogur S, Stankovic M, Scholz P, Arbel T, Bogunovic H, Bradley AP, Carass A, et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat Commun.* 2018;9(1):1–13. <https://doi.org/10.1038/s41467-018-07619-7>.
14. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G, Thrun S, Dean J. A guide to deep learning in healthcare. *Nat Med.* 2019;25(1):24–9. <https://doi.org/10.1038/s41591-018-0316-z>.
15. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. *Med Image Anal.* 2017;42:60–88. <https://doi.org/10.1016/j.media.2017.07.005>.
16. Maier A, Syben C, Lasser T, Riess C. A gentle introduction to deep learning in medical image processing. *Z Med Phys.* 2019;29(2):86–101. <https://doi.org/10.1016/j.zemedi.2018.12.003>.
17. Karpathy A. Convnetjs: deep learning in your browser (2014), 2014. <https://cs.stanford.edu/people/karpathy/convnetjs/>
18. Koenderink JJ. The structure of images. *Biol Cybern.* 1984;50:363–70.
19. Ringach DL. Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *J Neurophysiol.* 2002;88(1):455–63.
20. Grant Sanderson (3Blue1Brown). Taylor series – essence of calculus, chapter 11, 2017. Video: youtube.com/watch?v=3d6DsJIBzI4.
21. ter Haar Romeny BM. Front-end vision and multi-scale image analysis, volume 27 of Computational Imaging and Vision Series. Berlin: Springer; 2003. <https://doi.org/10.1007/978-1-4020-8840-7>.
22. ter Haar Romeny BM. A geometric model for the functional circuits of the visual front-end. In: Grandinetti L, Lippert T, Petkov N, editors. *Brain-Inspired Computing*, volume 8603 of Lecture Notes in Computer Science. Springer International Publishing; 2014. p. 35–50. https://doi.org/10.1007/978-3-319-12084-3_4.
23. Blakemore C, Cooper GF. Development of the brain depends on the visual environment. *Nature.* 228:477–8. <https://doi.org/10.1038/228477a0>. October 1970. Video: youtube.com/watch?v=QzkMo45pcUo
24. Doolittle B, MacLay E. *The forest has eyes.* Seymour, Connecticut: Greenwich Workshop Press; 1998.
25. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. *Insights Imaging.* 2018;9(4):611–29. <https://doi.org/10.1007/s13244-018-0639-9>.
26. VanVeen F. The Neural Network Zoo, 2016. <https://www.asimovinstitute.org/neural-network-zoo/>
27. Sutton RS, Barto AG. *Reinforcement learning: an introduction.* Cambridge, MA: MIT Press; 2018.
28. Narla A, Kuprel B, Sarin K, Novoa R, Ko J. Automated classification of skin lesions: from pixels to practice. *J Investig Dermatol.* 2018;138(10):2108–10. <https://doi.org/10.1016/j.jid.2018.06.175>.
29. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 2017;542(7639):115–8. <https://doi.org/10.1038/nature21056>.
30. Rothe R, Timofte R, Van Gool L. Deep expectation of real and apparent age from a single image without facial landmarks. *Int J Comput Vis.* 2018;126(2–4):144–57. <https://doi.org/10.1007/s11263-016-0940-3>.
31. Wolfram Research. Wolfram Neural Network Repository: Age-estimation-VGG-16-trained-on-IMDB-WIKI-and-Looking-at-People-Data, 2019. <https://resources.wolframcloud.com/NeuralNetRepository/>
32. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, Mahendiran T, Moraes G, Shamdas M, Kern C, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health.* 2019;1(6):e271–97. [https://doi.org/10.1016/S2589-7500\(19\)30123-2](https://doi.org/10.1016/S2589-7500(19)30123-2).
33. Diaz-Pinto A, Morales S, Naranjo V, Köhler T, Mossi JM, Navea A. CNNs for automatic glaucoma assessment using fundus images: an extensive validation. *Biomed Eng Online.* 2019;18(1):29. <https://doi.org/10.1186/s12938-019-0649-y>.
34. TensorFlow Model Garden, Google Inc. 2020. <https://github.com/tensorflow/models>
35. Caffe Model Zoo. Berkeley Artificial Intelligence Research (BAIR) lab. 2020. https://caffe.berkeleyvision.org/model_zoo.html
36. The Wolfram Neural Networks Repository. 2020. <https://resources.wolframcloud.com/NeuralNetRepository>
37. Luo JW, Chong JJR. Review of natural language processing in radiology. *Neuroimag Clin.* 2020;30(4):447–58.

38. Zhao H, Qi X, Shen X, Shi J, Jia J. ICnet for real-time semantic segmentation on high-resolution images. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018. p. 405–20. Video: youtube.com/watch?v=qWI9idsCuLQ.
39. Hesamian MH, Jia W, He X, Kennedy P. Deep learning techniques for medical image segmentation: achievements and challenges. *J Digit Imaging*. 2019;32(4): 582–96. <https://doi.org/10.1007/s10278-019-00227-x>.
40. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Lecture Notes in Computer Science, International Conference on Medical Image Computing and Computer-Assisted Intervention, vol. 9351. Cham: Springer; 2015. p. 234–41.
41. Moeskops P, Wolterink JM, van der Velden BHM, Gilhuijs KGA, Leiner T, Viergever MA, Isgum I. Deep learning for multi-task medical image segmentation in multiple modalities. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2016. p. 478–86. https://doi.org/10.1007/978-3-319-46723-8_55.
42. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. *J Pathol Inform*. 2016;7. <https://doi.org/10.4103/2153-3539.186902>.
43. Murphy K, Habib SS, Zaidi SMA, Khowaja S, Khan A, Melendez J, Scholten ET, Amad F, Schalekamp S, Verhagen M, et al. Computer aided detection of tuberculosis on chest radiographs: an evaluation of the CAD4TB v6 system. *Nat Sci Rep*. 2020;10(1):1–11. <https://doi.org/10.1038/s41598-020-62148-y>.
44. Wang J, Knol MJ, Tiulpin A, Dubost F, de Brujne M, Vernooij MW, Adams HHH, Ikram MA, Niessen WJ, Roschupkin GV. Gray matter age prediction as a biomarker for risk of dementia. *Proc Natl Acad Sci*. 2019;116(42):21213–8. <https://doi.org/10.1073/pnas.1902376116>.
45. Schlemper J, Caballero J, Hajnal JV, Price AN, Rueckert D. A deep cascade of convolutional neural networks for dynamic MR image reconstruction. *IEEE Trans Med Imaging*. 2017;37(2):491–503. <https://doi.org/10.1109/TMI.2017.2760978>.
46. Bahrami K, Shi F, Zong X, Shin HW, An H, Shen D. Reconstruction of 7T-like images from 3T MRI. *IEEE Trans Med Imaging*. 2016;35(9):2085–97. <https://doi.org/10.1109/TMI.2016.2549918>.
47. Chaudhari AS, Fang Z, Kogan F, Wood J, Stevens KJ, Gibbons EK, Lee JH, Gold GE, Brian A. Hargreaves. Super-resolution musculoskeletal MRI using deep learning. *Magn Reson Med*. 2018;80(5):2139–54. <https://doi.org/10.1002/mrm.27178>.
48. Nair LR, Subramaniam K, Prasannavenkatesan GKD. A review on multiple approaches to medical image retrieval system. In: Intelligent Computing in Engineering. Springer; 2020. p. 501–9.
49. Elharrouss O, Almaadeed N, AlMaadeed S, Akbari Y. Image inpainting: a review. *Neural Process Lett*. 2019;1–22. <https://doi.org/10.1007/s11063-019-10163-0>.
50. Higaki T, Nakamura Y, Tatsugami F, Nakaura T, Awai K. Improvement of image quality at CT and MRI using deep learning. *Jpn J Radiol*. 2019;37(1):73–80.
51. Yabo F, Yang L, Wang T, Curran WJ, Liu T, Yang X. Deep learning in medical image registration: a review. *Phys Med Biol*. 2020. <https://doi.org/10.1088/1361-6560/ab843e>.
52. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T. Analyzing and improving the image quality of StyleGAN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020. p. 8110–9. Video: youtube.com/watch?v=kSLJriaOumA&t=9s.
53. Wolterink JM, Dinkla AM, Savenije MHF, Seevinck PR, van den Berg CAT, Isgum I. Deep MR to CT synthesis using unpaired data. In: International Workshop on Simulation and Synthesis in Medical Imaging. Springer; 2017. p. 14–23. https://doi.org/10.1007/978-3-319-68127-6_2.
54. Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: a review. *Med Image Anal*. 2019;58:101552.
55. Zhu J-Y, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision; 2017. p. 2223–32.
56. Tarek Shaban M, Baur C, Navab N, Albarqouni S. Staingan: Stain style transfer for digital histological images. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE; 2019. p. 953–6. <https://doi.org/10.1109/ISBI.2019.8759152>.
57. Martin-Brualla R, Radwan N, Sajjadi MSM, Barron JT, Dosovitskiy A, Duckworth D. NeRF in the Wild: neural radiance fields for unconstrained photo collections. arXiv preprint arXiv:2008.02268, 2020. Video: youtube.com/watch?v=yPKIx0N2Vf0.
58. ter Haar Romeny BM. A deeper understanding of deep learning. In: Artificial Intelligence in medical imaging: opportunities, applications and risks. Cham: Springer; 2018. p. 25–38. <https://doi.org/10.1007/978-3-319-94878-2>.
59. Kandel ER, Schwartz JH, Jessell TM. Principles of neural science. 6th ed. New York, NY: McGraw-Hill; 2013. ISBN 9781259642234.
60. Kolb H. Roles of amacrine cells. In Webvision. The Organization of the Retina and Visual System. 2016. <http://webvision.med.utah.edu/>
61. Masland RH. The neuronal organization of the retina. *Neuron*. 2012;76(2):266–80. <https://doi.org/10.1016/j.neuron.2012.10.002>. <http://www.sciencedirect.com/science/article/pii/S0896627312008835>
62. Hubel DH. Eye, Brain and Vision, volume 22 of Scientific American Library. New York: Scientific American Press; 1988.
63. Levin A, Lischinski D, Weiss Y. Colorization using optimization. *ACM Trans Graph (TOG)*; 2004;23(3): 689–694. <https://doi.org/10.1145/1186562.1015780>.

-
- 64. Iizuka S, Simo-Serra E, Ishikawa H. Let there be color! Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. ACM Trans Graph (ToG). 2016;35(4):1–11. <https://doi.org/10.1145/2897824.2925974>.
 - 65. Seymour KJ, Williams MA, Rich AN. The representation of color across the human visual cortex: distinguishing chromatic signals contributing to object form versus surface color. Cereb Cortex. 2016;26(5):1997–2005. <https://doi.org/10.1093/cercor/bhv021>.
 - 66. Ninio J, Stevens KA. Variations on the Hermann grid: an extinction illusion. Perception. 2000;29(10):1209–17. <https://doi.org/10.1068/p2985>.
 - 67. Ghafoorian M, Karssemeijer N, Heskes T, van Uder IWM, de Leeuw FE, Marchiori E, van Ginneken B, Platel B. Non-uniform patch sampling with deep convolutional neural networks for white matter hyperintensity segmentation. In: 13th International Symposium on Biomedical Imaging (ISBI). IEEE; 2016. p. 1414–7. <https://doi.org/10.1109/ISBI.2016.7493532>.
 - 68. Nelson E. Wide-are indoor and outdoor real-time 3D SLAM, 2016. Movie: <https://www.youtube.com/watch?v=08GTGfNneCI>
 - 69. Ulrich Scheipers, Christian Perrey, Stefan Siebers, Christian Hansen, and Helmut Ermert. A tutorial on the use of ROC analysis for computer-aided diagnostic systems. Ultrasonic Imaging. 2005;27(3):181–198. <https://doi.org/10.1177/016173460502700304>.



Importance of AI in Medicine

6

Evolution, Processes, and Challenges

Katarina A. M. Gospic and Greg Passmore

Contents

Introduction	100
The Amazing World of AI	101
A World of Terminology	101
Neural Nets: Quite Similar to Neural Networks in the Human Brain	102
Data: The Fuel for AI	102
Segmentation: Feeding Quality Data	103
Training Mode	104
Multisensor Data	104
Prediction Mode	105
Levels of Expertise	105
Self-Evaluation Mode and Correlations	106
Breaking Boundary Conditions	107
Multiple Ways of Solving a Problem	107
Processing Improvements	109
Validation	109
Gatekeepers	110
Limitations	111
Responsible Artificial Intelligence	111
Various Agendas	111

K. A. M. Gospic (✉)
Brainbow Labs AB, Stockholm, Sweden

G. Passmore
VR Media Technology, Los Angeles, CA, USA
e-mail: greg@passmorevr.com

Transparency	112
When to Use AI in Medicine	113
References	113

Abstract

Artificial intelligence is a chaotic field with conflicting terminology, philosophies, and architectures. This occurs in part due to the rapid merging of practitioners in computer graphics, image processing, and computer science. The entry of large corporations, attempting to simplify and streamline machine learning, creates an environment providing for reduced technical understanding by users. Competition for noteworthy breakthroughs drive rapid prototyping and early reporting. This collision yields a vivid spectrum and hidden risk. This chapter provides an orientation to important components and terminology in machine learning, plus painless examples of success and failure.

Do not include reference citations or undefined abbreviations in abstracts since these are often read independently of the actual chapter and without access to the reference list.

Keywords

AI · Neural nets · Segmentation · Learning mode · Prediction mode · Self-evaluation mode · Perceptron · Forward propagation · Back propagation · Multisensory data · Boundary condition · Validation · Gatekeeper · Transparency · Artificial intelligence · Medicine · Cancer · Decision-making · Decision basis

Introduction

Would you like to know four years in advance that you have a high probability to fall ill to a potentially deadly disease? [13] AI technology is available today, for hundreds of medical applications, and is under vigorous investigation. How does AI work? Let us take breast cancer as an example

[15]. At its simplest, you first collect thousands of mammograms over a period of time. Then you mark which ones contain cancer and which ones do not. If you then go back to the negative mammograms, years before the cancer has occurred, and compare them to the mammograms identified as positively cancerous, the computer can search for patterns and possibly identify early indicators of the disease. Years before it happens. Once a new patient gets her mammogram taken, you ask the AI to compare her pictures with the thousands of examples already collected. Leveraging significant computational power, you can derive at an answer immediately, providing you with a probability score. This score indicates the likelihood of having breast cancer, or developing it in the future. The AI does this by comparing your mammogram with ones that look similar to yours. How many are marked positive for breast cancer, or, if they are negative, how many were diagnosed with cancer in the years after imaging? Depending on the probability score, you can determine a personal threshold for what will be considered at risk versus what will not. For example, during your screening, the displayed imagery of the mammogram could be color-coded as such: green if you are determined to be in the low-risk category, yellow if further investigation is recommended, and red if there is any indication of a high probability of breast cancer. In some medical specialities, this type of technology may already outperform doctors. This makes sense, as a doctor may see a few thousand patients throughout their career, whereas a computer could theoretically collect data from millions of patients undergoing mammograms. The beauty of this system is that every single patient, who contributes his/her data, is helping society as a whole. How well the AI machine performs is due in part to the algorithms, but also to the amount of meaningful data collected. It is important to understand that masses of data are the rocket fuel for AI.

This introduction is designed to give you, who are new to AI in medicine, a quick overview into the rapidly advancing world of artificial intelligence, machine learning, neural nets, and the unique hardware architectures being designed to quickly and precisely process these concepts. In this short read, we hope to bring you up to speed on the broad concepts, promises, and pitfalls of AI.

It will equip you to decide where to dig deeper and will allow you to read the upcoming chapters with a basic understanding of AI. With this knowledge, you will be able to create a map of where these application-focused authors are located in the technical space of AI.

Please note that this chapter focuses on machine learning for imagery, one of the most common uses of AI in medicine today. There are wide variations in techniques, nomenclature, and architectures, which is why we will take you on a simplified path, one without a lot of terminology or math. The path we are taking is not the only one, but rather a grand tour to provide some context.

The Amazing World of AI

The human eye cannot detect subtle patterns just by looking at a histopathological slice. But a computer can “see” what the human eye cannot [12]. Computers can find patterns, and are able to identify the slightest variations between images, on a pixel-by-pixel basis. This is the magic of computers. But how does AI actually work? How do we go from data input – we feed the computer images of breasts potentially containing cancerous cells – to data output, that is, the computer determining whether it is cancer or not?

At its most basic, AI can either be in training mode, prediction mode, or self-evaluation mode. Before we can get any meaningful output, we need to train the AI. The traditional way of doing this is by feeding in pictures, or other data, that we tag as cancer or not cancer. This is the AI training session. It is easy to think that the more pictures we feed in the better. But it is not as

simple as that. To create well-performing AI, we need to feed in pictures with rich variations and try to account for as much variation as possible. Usually, during training sessions, we tend to choose the pictures we are very sure of. The consequence of this gives us an AI model that computes a correct output for the “easy” cases. In order for the AI to perform well with more difficult cases, we need to train appropriately. A great challenge with AI training is to know that we have not added a feature that is present in all the sick cases, but irrelevant to the actual condition. A nonmedical example of this is a person who trained his AI to recognize horses. Most of the pictures he fed in were taken by the same photographer. What he did not notice was that the images containing a horse, also contained a copyright symbol with the photographer’s name on it. When testing his system with random pictures of a horse from the Internet, his system classified those pictures as “not horse,” because it lacked the copyright symbol and the photographer’s name. This being said, it is easy to miss such a small detail that is detrimental to the whole system. We need to remember what is obvious to us is not obvious to a computer. The computer does not understand causality, only correlation. Planning for training, and careful selection of data, is critical to success. Designing the AI model can also include setting limits or rules on the correlations to exclude results which lie outside of possibility. This may be accomplished through software to visualize correlation importance, to understand what the AI believes is important. Computers have no common sense, something we will circle back around to. AI learns only based on what we feed it.

A World of Terminology

When I was in medical school, in the early 2000s, I had a teacher that taught us the anatomy of a leg by using an old overhead projector from 1973. When we pointed that out, he replied that the leg anatomy had not changed since then. Terminology in the AI world changes constantly as AI is a rapidly evolving field. What is said one way today may be said differently tomorrow [2, 5, 10].

Nevertheless, it is important to understand basic concepts about the underlying processes of AI. Let us dive into the more technical part of this chapter to open the black box of AI. To do that, and to achieve an understanding of these underlying processes, we will start by defining some fundamental terminology.

Neural Nets: Quite Similar to Neural Networks in the Human Brain

There are different types of AI, and a common method to solve AI problems is through the use of neural nets. The purpose of neural nets is to imitate human intelligence to solve a problem. The core feature of neural nets is that it helps to correlate or declutter information. Associated information is enhanced and enables us to focus on it, whereas less important information is given less attention. We can think of this as circuits in the human brain. Neurons that fire together, wire together, creates a self-regulating system and contributes to noise reduction. So as input information is more important, it is given more weight.

Similar to the human brain, the output from an AI neural net is dependent on the strengths and importance of stimuli. Visual input through the human eye activates different areas of the visual cortex. Computer vision, which can use the architecture of neural nets, works in a similar manner. Computer vision consists of highly specialized neural net circuitry that processes a continuous flow of data, typically with circuits for edges, stereo fusion, and contrast management, just like specialized areas in the human visual cortex. So, when the human eye sees a cat running over the street, this picture is also segmented in order for different parts of the brain to focus on motion, edges, colors, background reference, etc. The computer also segments the picture, as we will see soon, and may also contain specialized region processing for these different features.

The core unit of a neural net is a perceptron, if one has a medical background I think it helps to think of a perceptron being a neuron. The purpose with the perceptron is to weight the importance of incoming variables (input) and generate an output.

This is called forward propagation. Unlike typical statistics, these circuits handle huge numbers of input parameters and are evaluated in something called the hidden layer. The hidden layer is located between the input and output of the process and performs the actual weighting. The hidden layers typically use nonlinear weighting, for help force a decision. Weighting individual perceptrons is the mathematical process of increasing or decreasing the importance of each specific node. This process distinguishes perceptrons from typical statistics.

Single perceptrons can only retain a tiny amount of information. Single layers or perceptrons have limits in the degree of convergence that can be obtained. In contrast, multilayer perceptrons also called deep learning, with many layers, have greater convergence power and can take more complexity into account. So, when we later on talk about data segmentation, data fusion, and data federation – these are all processes that tend to imply multiple layers of perceptrons. The term deep learning refers to this concept of numerous stacked layers of perceptrons, so that each perceptron layer can focus on a subset of the overall problem or a degree of refinement. Deep learning is trusted to perform more advanced AI in medicine, which makes sense considering the subtleness of a lot of medical diagnostic imaging. The example in the coming segmenting section is a real-life example of deep learning.

Data: The Fuel for AI

For AI to work, we need to convert an image into numbers which can be feed into the machine for learning equations, which we will describe below. At its simplest, images can be represented as a simple matrix of numbers. Groups of images may be more conveniently represented as tensors. Tensors are special spaces for numbers, which are often multidimensional. AI input can vary in modality as well as dimensions. It may, for example, be a single image or a sequence of images either in space (such as a 3D biopsy image stack) or in time (such as a movie). A different modality may also be spectral in nature, e.g., be 600 nm

wavelength imagery or 800 nm. For illustrative purposes we will talk primarily about images in the coming sections. The example we have chosen is *one* way of doing AI. This example revolves around an experimental multimodal imaging system developed by one of the authors [Passmore]. The device, let us call it HyperTrak, was developed for noncontact imaging of people to develop a health score, based on detected symptoms of illness. The motivation was to obtain some clue to who are most symptomatic of illness, and are possible candidates for further screening. The driving factor for health scoring is the ongoing crisis of covid-19, especially in large crowds. This provided an opportunity to leverage AI, develop a hyperspectral imaging workbench, and make a contribution. The device has been built and is currently being tested in Texas, USA. It is also a convenient platform to reference in these discussions.

Segmentation: Feeding Quality Data

A traditional way of teaching AI is to do it in binary form, as in “sick” or “not sick,” while looking at a picture as a whole. But there can be so much more to it than that. Let us describe the process of feeding more complex data into neural nets and how deep learning works, in very simplified terms. The HyperTrak scanner consists of different sensors which pick up different aspects of light. The captured images typically not only contain the color of interest, but also texture, edges, geometric properties, background clutter, and more. All of these properties are different variables. When extracting portions of an image containing different properties, we call that segmentation. Looking at individual image segments, we can tease out commonalities in that region based on variables, such as green regions, areas with a bumpy texture, specific cell types, and so on.

Imagine a crowd of people passing through the entrance of an office building, and the HyperTrak scanner is placed in front of the entrance so that everyone who enters the building will pass it. The first thing the scanner captures is a crowd at the

entrance doors along with all the things in the background. Segmentation is the process that separates the people from the background. In a second segmentation process, individual faces are separated from the crowd. A third segmentation process starts to pick each face apart into different types of facial regions. Think of this process as picking Mr. Potato Head’s head apart. You can remove his eyes, nose, ears, and mouth. One important facial region that is scrutinized in this context is the area under the eye (AUE). The AUE is normalized for light and color before it is picked apart into its different variables, e.g., sweat, temperature, pallor, etc. Each of these variables are then fed into a set of neural nets which consists of neurons (think of it as a box for AI), synonym with perceptrons. By normalizing the data and sorting it by different variables, the dataset becomes better “behaved,” making it easier to understand which image segments and which variables contribute to the output. Imagine we segment the picture of a person into ten unique layers, or in our example ten different parts from Mr. Potato Head’s head. Now we have ten images, with isolated objects, to look at, rather than just one (i.e., the complete picture of a whole Mr. Potato Head).

When we feed the temperature properties of the AUE into a box of neural nets, we get a scatterplot of the temperature data. A polynomial equation is fitted to the data which is a large-scale operation. Numerous equations (layers of neurons) are stacked on top of one another in order to diminish the error factor and finding the best fit for the curve. Solving the equations results in a line that is fitted to the data. The reason for layers upon layers of equations is that each pixel is connected to all neurons in the neighboring layer, and each of those neurons are connected to each neuron in the next layer in line, and so on. The purpose of this rather complex structure is to drive the error toward zero to get the best possible fit to the data.

The amount of equations contributes to complexity of scale, requiring substantial computational power. Once we have the fit in training mode, we will tell the AI: “when the line looks like this, it is covid.” In prediction mode it will be: “when the incoming line looks like this it falls into

the covid category.” The text above only described the process of analyzing one variable for the AUE. Now imagine we have other regions of the face that we are interested in, e.g., the actual eyes, forehead, etc. In the same way that Mr. Potato Head has different objects we can pick apart, we can perform this procedure for all the areas/objects we are interested in. The more numerous our sources of input data, and the more variables we put into the AI from our segmented objects, the more information we have to base our output on. This is a type of pattern recognition via machine learning. Pattern recognition is a widely used term that indicates when computers learn to identify patterns. Summarizing, segmenting the data into different layers provides distinct channels about what “sick” looks like, compared to just having a single layer (the whole picture). It also makes it easier for us to tease out the contributing correlations, assisting us in appropriate supervision.

Training Mode

Training can occur in different ways. One way of thinking is that when feeding the AI, we can think of it as providing a verb, a noun, and an adjective. Let us take an example of a scanned image, positive for covid and a resulting patient death. This scenario contains more than a noun (covid or no covid). Please think about this for a moment. Compare the above sentence with a scanned image, positive for covid and the patient was healthy within a week. The last sentence adds the information about a good outcome compared to the first sentence. The difference may be the degree of illness, the adjective. The predicted existed of covid, the noun. The death a verb. If AI indicated covid – would not you also like to know the last part of the sentence, the end result? If yes, this is something we need to add into the AI training, so this message can be included to you. Thus, it is highly desirable when we feed AI, to give it complete information, rather than just label input as sick or not sick. It is also desirable for us to show variation in cases. Going back to the example with the horse. Imagine that we only

feed the AI with pictures of horses standing still. Then the AI may perform poorly at recognizing horses when jumping, running, or lying down. Or even a pony. It is similar in medicine. If we only train with covid from one vantage point or with a single type of people, but not with another, the computer will have no reference when we provide new and unfamiliar data. As such, it will not be able to reliably detect if the result is sick or not sick. For this reason, we need to plan and strategize for our AI, before training and deployment.

Multisensor Data

Let us revisit multimodality. A powerful option of data input into AI is the addition of multisensor channels. When we want a complete picture of health state, we often do this best by collecting a variety of data types. In the introduction, we presented how one type of data can be utilized to foresee disease. For example, mammogram data to predict breast cancer. However, one powerful technique in AI lies in the combination of different data types. So, if mammography can be considered as a single channel of data, adding information from an additional type, such as ultrasound, can be considered as a second channel of data. Adding different channels of data is called multimodality. You can imagine one example of multimodal data collection by thinking of the five senses of the body. Imagine you are invited to a fancy dinner at a restaurant located by the sea, where you are sitting in a hanging garden of lemon trees. If you only collect data from your eyes, you will have one type of experience. If you collect data from both your eyes and your nose, you will have another type experience. And if you have your eyes, nose, and ears there will be a third experience. The more senses you use, the more complete your perception of the dinner will be. Of course, forgetting the sense of taste would be a disaster. It works the same way within medicine.

Data collected from different modalities can be normalized and placed in the same space. A simple example could be cameras with different resolutions. It simplifies our thinking to overlay the different cameras in a common resolution. This

process is called data fusion. In addition, we may want our AI module to communicate with other AI modules. These other modules may be other modalities or it could be something as distant as another doctor's AI for a related disease. We do this with federation, when different AI modules communicate with each other, sometimes over great distance. When properly staged, the resulting answers can be the result of a collective of AI modules working together. As previously mentioned, these processes require multiple-layers perceptrons. These concepts are an emerging area of AI and hold tremendous promise.

Prediction Mode

So far, we have trained our machine for what is probably ill (check for covid) and what is probably not ill (lower priority to check for covid). Now, we want to see some predictions. To get a prediction, we need to switch the AI from learning mode to prediction mode. Imagine we now take a random picture of a person and feed it into the system. How can the AI know if it is covid positive or negative? What happens in our example is that the picture is segmented into Mr. Potato Head's different parts and their variables. For each segment we will use the matrices of weights and bias values. The computer applies the matrices to each segmented image's variables to determine the confidence of covid, as individual values from each segmentation. Let us say that all our sums, for all segments, fall into the positive sick/covid category, then the AI will give a positive "sick/covid" probability as an output. It could be that if only three of our curves are within the covid threshold, we may choose to assume "no significant symptoms" as an output. But what about if five or six of our curves fall into the positive covid category – what will the output be then? Frequently, the outputs from individual channels are fed into additional layers of machine learning. However, unless the answer is overwhelmingly distinct, we will ultimately have to make a judgement call regarding our belief whether the result is significant or indeterminate. This highlights the importance of determining thresholds and may

also, over time, help us understand the stability and usefulness of the predictions. One possible factor to add is the input confidence used to train the machine. If we only teach the AI that the output can either be covid or not covid, then prediction stability may suffer since the training is based on absolutes, where absolutes probably do not actually exist. However, if the AI is trained with tagging that indicates a 20% likelihood of covid, a 60% likelihood of covid, or an 80% likelihood of covid, it may be possible to avoid some errors of oversimplification. We need to ask ourselves, what is the basis for decision-making that we want to use? Thresholding, already an existing problem in diagnosis, carries through in machine learning. One can, however, make matters worse by oversimplifying data tagging used during training. We may also want different outputs for different indicated diseases depending on their severity. If we change the word "covid" from the example above, to "cold" or "cancer," the importance of associated thresholding and the importance of individual symptoms will vary.

We also need visibility into how predictions are weighted. Common sense tells us that some factors are of greater importance for indicating a disease than others. Meaning that even when only three modalities of an AI model indicate "covid," there may be a greater chance of the disease than if compared to six less important modalities indicating "covid."

Levels of Expertise

It is not uncommon to develop multiple models for a problem. This is different than modalities, in that a model is an architecture, or design of an AI solution. A modality is a characteristic of an input channel, which can be run through many different models. We can see the simultaneous use of different AI models as an analogy to a committee of doctors who are voting whether your data indicates covid indicators or not. If you only look at one model, it is a bit like asking a single doctor if it is covid or not. If you look at many models, let us say several – it is like asking several doctors if it is covid or not. However, it is also a similar thought

exercise as if those several doctors just recently finished medical school or if instead you have several highly experienced specialists in your committee. We all know there are some doctors who are more skilled than others. When the voting concerning your health is happening, you want the best doctors on your committee. Especially when this voting concerns potentially deadly illnesses such as covid or cancer. For this to happen we need to consider all the above. The above reasoning also stresses that we need to be humble about AI and possibly seek the opinion of multiple models. We unintentionally transfer our bias and errors into AI training. Being poorly skilled or being genius, our opinions will carry into the AI model, which means in some ways we are replicating ourselves with AI. Therefore, we need to be humble about ourselves as well. The human brain created silicon intelligence and subsequent ramifications. If the trainers are mistaken about reality, we can expect our AI predictions to reflect this.

AI can perform tirelessly, in terms of time and energy, as well as process huge amounts of medical data. In the real world, a human doctor may go through a few histopathological slices from a cancerous tumor to determine diagnosis, stage, and prognosis. A computer is able to go through thousands of slices, several times. Similarly, it could screen a crowd of people for covid infection in a second. This enables us to be more thorough in the medical care provided. It may also increase the probability to be more accurate, while the chance of missing important information decreases. It is predicted that AI will have the greatest immediate impact on specialities such as pathology and radiology, because these are specialities based on visual data.

At the moment, AI is being celebrated. It is similar to when we start dating. At first we believe the number of desirable partners is great. However, after a while we probably notice that 90% are not compatible and fit into the dates that we should discard. It is the same with AI, not all AI is a good fit. Most early experimentation may need to be discarded. The consequences of having a bad date are typically inconvenience, but the consequences of poorly trained AI in medicine can be fatal. Level of expertise is crucial. With

that said, we need to seriously consider who we date, and in medicine, do a proper background check.

Self-Evaluation Mode and Correlations

Once we have taught the AI system what covid, cancer, or any other disease looks like (or not), as well as having informed the system which variables are more important than others, it is time for the system to learn to improve itself. It does this when in self-evaluating mode. Perceptrons can refer to themselves in feedback loops, enabling the system to self-evaluate for the purpose of convergence.

This happens when the actual error factor coming out of the hidden layer is compared with the expected value. Gradient descent is one process of minimizing the error and updating the weight for prior circuits. As previously mentioned, imagine that every variable in the neural net is a person in a committee who will vote if a person has covid or not. If the discrepancy between the expected error and the actual error is huge, it means that this person (perceptron) is bad at predicting who has covid or not. When there is a vote, this person gets it wrong most of the time. Such a person would be considered less important in a committee. We would give their opinion a lower weight, possibly much lower. In contrast, other variables that are very performative, these are considered to be very competent, similar to a great doctor. Such a perceptron will be given considerable weight in the voting process and, subsequently, contributes more to the output. This self-evaluating feedback processes is called back propagation and was a critical breakthrough between rule-based AI and self-learning machines.

Related to correlations is the ability of AI to pick up changes in correlations. This is valuable when the symptomatology of a disease changes. Using such a system could be perfect to detect, for example, how a covid mutation may cause changing symptomology of the disease. Such a method could give early alerts to authorities that the symptoms are in motion, which could prompt investigation. For example, imagine that the variable

which is the best predictor of covid today is pallor, but after a new mutation has taken over, we may see that this variable is contributing to more errors than usual. This could be a sign that something is happening. In parallel, another variable, let us say temperature starts becoming a better predictor. The system could then allow for the perceptron weights to change. In this scenario, the pallor variable could switch from having a great weight, or importance, to become much less important and the opposite could happen for the temperature variable.

Computers are excellent at picking up correlations. However, they are not strong at understanding if a correlation is meaningful. When AI gives us information about how two or more variables may correlate, it is up to us to understand causation and create the narrative around such a finding. In the horse example the copyright symbol was right every time, in identifying a horse, until the input data was taken from another photographer. With this said, we as humans need to understand causation and create the narrative around relationships. Here, human supervision can add sanity to the process.

One interesting process is to ask the AI to create its own image from a tag, such as “sick.” This process of generating what the machine thinks, or internally sees, can be instructive. Under supervision, images which appear correct can be fed back into the learning cycle for reinforcement training.

Breaking Boundary Conditions

Again, AI is dependent on what it has been taught. For example, say that all images of people that have been feed into HyperTrack had both brown eyes and were sick. Meaning that by poor design or accident, it could be assumed that covid-positive people have brown eyes. If HyperTrack would be used in, e.g., Sweden, where lots of people have blue eyes, the system may not reliably predict until trained with Swedish people. We refer to this as breaking a boundary condition.

There are nonintuitive stories around this. In one case an AI was trained by letting young men

select and add pictures of males and females into the computer, teaching it whether the picture showed a male or a female. All female images selected by the young men were females with large breasts. When testing the prediction capacity of the system, females with smaller breasts were also added and predictably, these pictures were misclassified as men. Other examples of these AI training mishaps have occurred with concerning social consequences, such as racial bias based on skin color.

It is also helpful to let the computer know if we are teaching it several things simultaneously. If we put in multiple covid symptoms, but do not differentiate in our architecture, we can expect slower convergence, or unexpected results. If we do not, the computer will not know that we taught it two different things and teasing apart the weighting of these different symptoms can become more challenging. Purely as illustration, we have included a picture below of a “dog-flower.” In this case the AI is unaware that a dog is not the same thing as a flower. Although clear to us this is nonsensical, although interesting, it highlights that one cannot expect AI to possess common sense. We would like to use this image to help re-emphasize that the computer will only learn what we teach it. Therefore, it is important to consider what type of biases and omissions that may exist in the data, in order to derive accurate predictions from widely varying data and possibly intertwined case modality. We also need to train and update AI systems continuously to stay current, as the world around us changes.

Multiple Ways of Solving a Problem

A proper background check in AI is about understanding the ways a problem can be solved. When we understand the basic concepts we can start to evaluate model validity. We also begin to experience humility for all the things that can go wrong. There are many possible approaches to problem-solving in AI. There are a wide variety of architectures, preexisting models, and differing hardware platforms. These selections can have dramatic impact on success and practicality. Let

us look briefly at one version of recent AI history as it is related to image-based learning.

What most of us today think of as AI is considered machine learning. In our example, we are teaching a computer what is possibly covid or not, from imagery. AI is the parent of machine learning and a lot of early work was performed using rules, not self-teaching. Today, most machine learning is not built so much on explicit rule definition. Some of AI's roots are reflected in an early AI language, Lisp, and an early computer built around it, Symbolics. Symbolics popularized rule-based algorithms, such as flocking. This programming philosophy, which can be thought of as object personality structures, was applied to model natural phenomena, including particles, fractals, and eventually chaos. These rule-based methods became what were called expert system. This hardware manufacturer, and others like them, made inroads into the applied sciences of computer graphics and image processing.

The computer graphics group focused on creating models to create images of the world. Although math was certainly a large part of what was called AI at the time, the math was explicitly defined and did not contain self-teaching. It did use stochastics [3, 11], a form of randomness, to insert a sense of natural complexity. This created a sense of machine intelligence. Watching computer animation of flocking certainly gives an impression of machine intelligence, where there is no real opportunity for self-modification. This community was focused on the output of imagery and was most visible in visual effects for motion pictures, computer gaming, and computer art.

At the same time, another community, who had largely separate conferences and publications, was at work on pattern recognition and image processing. A group of these image processing programmers were focused machine vision, the practical understanding of input imagery. Machine vision programmers looked at a very similar problem to the computer graphics tribe, but in the opposite direction. How to take real-world imagery and create an internal model. This group was most visible through aerial mapping and factory automation.

A third group, composed more of physicists and mathematicians, had been working for decades on the overarching mechanics of AI. To an outsider, it appeared they were more focused on the theory of AI. Suffering numerous setbacks, a series of impressive breakthroughs regarding methods of self-teaching began having some serious success in the 1990s. This group gained attention through AI gaming of chess and robotics.

Today, the three groups have largely joined, each bringing their own vocabulary, methods, and architectures with them. This, in part, explains the diversity of terminology and methods used in AI. Every group have an ownership stake in AI, while simultaneously approaching it from different perspectives. Explicit operations, symbolic AI, and rule definition have largely given way to machine learning, a method of automatically converging on pattern recognition. The clarity of rule-based AI then gave way to convergence and abstraction.

Machine learning (ML) also shares commonality with statistical models [1]. Both use data-driven model generation, both are designed to attempt understanding. ML, however, is almost solely focused on prediction, where statistics, at least in theory, is focused on teasing out causality. ML produces these predictions through convergence, without necessarily yielding an explicit understanding of the relationship between variables. It is the drive for prediction that uniquely defines ML and holds much of its challenge. Those working in the field see remarkable success and equally remarkable failures of ML, frequently resulting from relatively minor processing adjustments or inadequate verification. It is as if a thinking machine is willing to go way out on the limb, in its aggressive attempt to predict. When successful, ML does not typically serve up knowledge of the relationships between input variables. We then, as operators, may blindly accept the predictions, based on feedback loops and testing, with no real understanding of the underpinning relationships. This is becoming increasingly common as scientists shift toward data feeding of more abstracted systems, without any involvement in the programming or algorithm development. ML learning then can be thought of as a form of

consequentialism whereas statistics is rooted in deontology. Consequentialism is deeply rooted in programming culture and it seems only natural that the implementation of ML systems has evolved from this mode of thinking. After all, massive computational loads are a common form of permissible harm in programming, only to be minimized later when the end result is successfully achieved.

This in no way diminishes the importance of ML, but provides the context for the radical shifts and complexities that occurred leading to today's popular definition of the term AI. Later in the text we will therefore discuss aspects to consider, as we increasingly lose an understanding of how the machine thinks, in order to provide classification and prediction.

Processing Improvements

When I was working with imaging processing during the first decade of the 2000s I always heard fun stories from elderly colleagues about how they had dozens of pictures in the 1980s and how they had to cut, paste, and glue them together to make sense of the data. When I was doing my scans, I got thousands of pictures from every scanning session. I did not use scissors or glue. I used a software called Matlab. In Matlab there were several pages of code that helped me automate the preprocessing of my imaging data. A single error in the code, like a comma or a typo, could result in fatal consequences for the analyses. An AI may traverse a one million lines of code! In the same sense that several typos have been made while writing this text, typos are made when writing one million lines of code. From previous paragraphs we have learned that thousands of equations need to be solved to weight the importance of different aspects of data. With this said, AI requires tremendous computational power and there are lots of things that can go wrong.

Committing a mistake within research leads to false results. Depending on what research it is, the impact it has on patients may vary. A nonsolid AI system can give the wrong output. Not just to a single patient or a few patients, which would be

the case if a doctor committed mistakes. It could be amplified to thousands of patients. What does it mean when thousands of patients suddenly get the information that they have a high likelihood of having cancer? This means that we need to think when AI gives us a prediction. We need to discuss what to do with the information. Imagine you get an output recommending giving the patient in front of you cancer treatment. Do you really want to give them radiation, chemotherapy, or amputate their leg? This is a decision you have to make on solid ground.

Validation

This brings us to the topic of validity of AI [3]. When AI is created, it can either be made from scratch or we can stand on the shoulders of giants using different AI prepackages to create our tailor-made AI system. How do we know that the pre-made AI packages are good? We have learned that a horse picture without a copyright symbol can be classified as not showing a horse and women with small breasts can be classified as being men, because of their small bust size. A recent anecdote of encoded bias has been on display as AI tagged members of the US Congress as known criminals, more often identified minorities as a credit risk and an abject inability for everyone to understand how an AI model makes decisions. The confidence interval for the AI systems predicting US Congress members as criminals was 80%. Eighty percent sounds pretty good, but less so if you are identified as being a convicted criminal (even though you are not) due to skin color or if you get a cancer diagnosis and go through a cancer treatment without having cancer.

Evidently, a crucial topic to discuss around AI is how we validate the system. From a basic medical research perspective, we always talk about the importance of replication to validate a result. This becomes tricky with AI because you can only get an identical result from the system if you feed it identical data. So how do we do it then? There are different ways one typically uses. One common option is to build two AI systems – each based on a different AI model.

When data is fed through and both systems give similar results, it would help validate their output. Another method almost always used is to test the system by feeding it data where you know the output, for example, “no cancer” or nonsense data. If the system manages to give you correct answers in both cases, it strengthens the confidence in the system. Doctors still need to be smart when handed an AI output; we cannot blindly trust the output as its processes can contribute to doctors losing their skills. Doctors should become more alert and smarter to keep up with the machines, so we can program them to be even better. Every output needs to be validated whether it is plausible or not.

As we have previously seen, computers are extremely good at finding correlations, but you as a clinician need to know your physiology and pathophysiology to understand if the result makes sense or not. In a medical context this means that if two different cell types in a histopathological slide covary, you as a doctor need to figure out why. As in a hypothetical case with pancreatic cancer, the presence of red blood cells in a sample may be important, as it may indicate a more aggressive cancer with greater vascularization and thus greater risk for bleeding, or it is a nonsense result that is not relevant at all for the condition. AI points our attention to what may be of importance, but so far, we need to create the story around it and understand the causality.

From an ethical point, we also need to decide the importance of AI. What happens if a doctor contradicts the AI output or vice versa? This maybe be extra important to consider in a country where an AI software produced in a fancy tech hub, has a greater status than a local health worker in a poor country. What if technology tells us to do the wrong thing, where a human eye can tell it is obviously wrong? Like the horse problem. We can be pretty sure that a picture shows a horse even though it does not have a copyright symbol. A security feature could be that humans can override the tech so it will not be forced to do the wrong thing. However, this will at the same time leave a loop hole for the ones who think they are right, but are actually wrong. The importance of an output will probably be in relation to resources. In a

country with restricted medical resources, the benefit of relying fully on AI may be the most beneficial option compared to no health care. To figure this out, a risk and benefit analysis would need to be in place. The community in which an AI operates needs to have guidelines for what weight to give AI output. In America these questions are put to the test as lawsuits are coming in against AI. Who is responsible for an AI output? Is it the company creating the AI system, the programmer, or the doctor using the system? Or could it be the hospital providing the AI tool to its doctors? Similar questions rise for autonomous vehicles, what happens if a self-driving vehicle runs over a child, who is responsible?

Gatekeepers

In this chapter it has become evident that AI in medicine is a complex topic and a very powerful tool. We need to decide who will be the gatekeeper of deciding what type of data is of good enough quality and certainty to be fed into a system, as well as what level of knowledge do you need to have to create AI for medicine.

What is the optimal process when creating new algorithms for medicine? This is also a core question within AI in medicine that should be thoroughly discussed within the profession. The most important is to know the individual AI system's strengths and limitations, especially the latter. By understanding what different aspects a system considers in combination with what type of data has trained the system, one can get a picture of the system's reliability. For more advanced diagnostics and where there is a huge risk for the patient, one could consider multiple AI systems verifying each other in order to be safe. One could consider a similar mechanism for AIM as for verification of blocks in a block chain. The greater the transaction, the more verifications are required for the transaction to go through and be valid. Using AI in medicine could rely on the same principle, the more of an impact an output has, the more validation there needs to be.

We also need to know or even decide who should train the AI. From a performance

perspective it would be optimal if the best experts in a field trained the AI that would be used within their area of medicine. But is it really the best of the best that trains AI today? If not we need to have a discussion who is. If it is an average Joe who has trained an AI system it means very different things if we read that the system is as good as doctors or if it is better than doctors. It is also important to know the level of expertise for those doctors. On the other hand which doctor is prepared to admit that AI is performing better diagnosis and prognosis than they do? When doctors do, they will probably be replaced. Meaning we wish for an amazing AI system that somehow is better than us, but we do not want to be replaced.

So far, AI seems to be at best, as good as doctors. Thus, at present, the recommendation is that AI should collaborate with doctors. This brings us to the topic of limitations.

Limitations

As AI relies on its data, programming, and algorithms, the design of these become crucial. So far, AI can never become better than the human programming it, thus imperfect humans code imperfect machines. In an AI context this means that if we use, for example, a mammogram where a doctor wrongfully labeled it as cancer (even though it was not), we will teach the AI the wrong thing. We may also tend to feed the AI with cases we are sure of. This gives us an AI that is good at solving the easy cases but has a higher likelihood of committing mistakes when handling ambiguous cases. Often we make the mistake of comparing humans versus computers in order to announce a winner. There is a danger of doing this. Computers as we know have amazing features, but so do skillful humans. Having common sense and feelings are actually good tools for decision-making, especially when balanced with concrete rational data. Going toward one of the extremes usually results in poor decision-making. To further understand the threats around AI and relying too much on rational data – let us compare AI to a sociopath.

Responsible Artificial Intelligence

A sociopath may have all the answers to curing your disease, but be insensitive to your emotions, make treatment decisions based on your social value, and may gloss over the specifics of your unique case. Your concerns about ethics, morality, and the pleadings for your life go in deaf ears. The eyes of the sociopath are distant, unblinking, and laser focused. Due to raw brilliance, should we all fall to the feet and obey this sociopath?

Today, thousands of us are building the most brilliant sociopath of all. Able to avert, diagnose, and treat disease, plus predict your mortality, many of us relinquish control of our lives and fail to understand the basis of decisions that will impact every future generation of humans. We are at a crossroad between a future of subservience to a silicon overlord and a future of a profound collective intelligence to cooperatively guide our world. If we close our eyes, blindly trust our sociopath friend, and run toward the cliffs, we will certainly get what we deserve. A lot of engineers hate drama and many doctors are simply numb to it. Drama, however, emphasizes a potential truth. It tunes our senses and energizes us into diligence and action. Surrounded by advanced technical lingo, let us pause and find a way to incorporate compassion, morality, and even humanity into these technical processes.

Various Agendas

The goals of AIM are filled with honor and pride. Collective intelligence, consistency in decision-making, sensing the hidden variables, and error checking live health providers, AI holds the promise to determine risk of disease onset, predict treatment success, manage complications, assist in ongoing care, and provide valuable advice on drug development and clinical research. All these goals appear within reach, with more powerful computers, vast storage farms, and absolute mountains of data.

Simultaneous to this, there is a small, but growing culture of technologists desiring what is a trinity of fairness, accountability, and

transparency. Why are these needed? What contributing factors require this mode of thinking? We see three main categories for diligence; AI's impact on doctors, on patients, and with regard to society. Let us briefly review these.

AI can lead to a de-emphasis on human agency. A key factor here is a human bias to trust automation results. Add to this, less time for doctors to interact with patients, due to schedule overloading, growing regulation on caregiving flexibility, growing legal liabilities, and administrative demands for doctors to be data entry clerks. As doctors become increasingly distracted by operating computers and following rules, they also become dissociated from their patients and isolated from the greater medical community. Patients, sensing they are on an assembly line, increasingly seek answers in social media and the Internet. Search engines, which are optimized to tell users what they want to hear, dish up biased results specially tailored for them. Differences between advice from their medical practitioner and what they see online lead to increasing mistrust of the medical community. This is certainly evident in the USA, where conspiracy theories are causing substantial social disruption. Patients, relying on information found online, can then justify unhealthy decisions regarding behavior and may be lulled into inaction due to errors of omission.

There are also broader forces at work. There is competition between the use of medical AI to save money versus saving people, a blindness to localized cultural values and a growing tendency to determine success based on patient satisfaction rather than treatment outcome. Encoded bias has been mentioned in earlier paragraphs.

On the broadest scale, there are issues surrounding the ownership of medical models and the enticement of large corporations to offer free or discounted AI services in exchange for AI data being under their control. Do these corporations have our best interests as a priority? There is also the risk of medical AI manipulation for use in cyber warfare, against both military and civilian targets. We need to consider the consequences of a foreign nation hacking into our systems, modifying the results of patients. Perhaps you would

receive an altered diagnosis, prognosis, or treatment. These issues illustrate the importance of human medical supervision, monitoring the plausibility of an AI result [13, 15].

Regarding AI, there are more fundamental risks, such as the risk of predictive inaccuracy. This risk may be due to generalization, unstable correlations, AI cheating, incorrect thresholding, overfitting, and lack of clinical context. These issues arise due to errors in medical records, multisyndrome patients, biased sampling, incomplete data, uncertainty in cases, and the error of data crowdsourcing instead of expert sourcing. Conflicting data can lead to unstable predictions, where a few new cases may cause dramatic shifts to outcome.

Transparency

Of course all these problems are correctable with diligence and imagination. The desirable path requires a few things: transparency, ubiquitous mobile sensors, federated thinking, and patient involvement [6–8].

An emerging field in AI is thinking visualization [9]. Hidden layers, especially of the stacks used in deep learning, is a black box to users. Akin to polynomials on top of polynomials at a large scale, these can become completely opaque to understanding. Methods of viewing the hidden layer connections, the priorities of input data attributes, and the linkage between correlations offer a key to thinking evaluation. The use of massive, coded point clouds, optimized for real-time display, can illustrate how a result was reached. As deep learning models become increasingly massive, our ability to render coherent abstractions of thought will be critical to transparency and detection of boundary violations, such as, not all hair is black, brown, blonde, or grey. Other methods of transparency are in the works as well, although visualization seems to have some of the densest data content.

As long as humans are encoding data, humans will be the bottleneck. Eventually, sensors with legs will be needed to feed the machine. The ability for an AI engine to gather its own data

and travel to the locations where such data is available will shift the equation. Robotics with a mind-numbing number of sensor modalities seems an inevitable future. Trained robotic observers that can see the visible, as well as the invisible, can begin to maintain the feedback loops of treatment, outcome, complications, mortality, litigation, cost, happiness, and predictive confidence.

At some point, as convergence occurs in specialties, multisyndrome complications will become an increasingly significant variable. Federated systems linking arrays of AI models will emerge, and ontology, coupled with communication standards, is a natural path. Today, competing companies are motivated to tunnel users into their farms and sell knowledge to financially vested third parties. Privacy regulation, ironically, can justify data silos to the benefit of large corporations. The development of open standards for AI models and their federation will be critical to averting monopolistic behavior. This flies in the face of open market financing, but hopefully medical data may eventually become an exception to data hoarding and alternative ways of funding AI research need to emerge.

When to Use AI in Medicine

A risk and benefit analysis should always be considered when implementing a new system to understand the consequences of a methodology. In the case of anemia it could be of great value to automate diagnostics and treatment as the procedure is straightforward. For example, blood tests indicative of iron anemia lead to a computer recommending you iron supplements. The risk with this procedure is fairly low. Even though you could be misdiagnosed, the consequence of such wrongdoing is negligible. Diagnosis of iron insufficiency stands in stark contrast to getting a cancer diagnosis where invasive treatments may be recommended. However, one may prefer an AI doctor to nothing. In rich countries we may have the opportunity to choose, whereas in poorer countries AI doctors may be the only ones available.

In a perfect world, vetted AI could make advanced health care available to the masses. However, before implementing a new system the entire value chain needs to be scrutinized. It is one thing to provide a patient with a diagnosis or prognosis, but what will the patient do with that information? A system needs to be in place to take care of the patient once the patient has been provided the results. Not all countries have straightforward systems to handle the medical care process from a to z. Thus, before implementing AI, there need to be clear structures and procedures regarding what the care process will look like. The patient needs to be able to act upon the provided information. Otherwise, one should consider the value of providing such info, if the patient is unable to act upon it. This is an ethical discussion. In some cases, it may be valuable for a patient (and/or their loved ones) to know that they suffer from bipolar disease, as this knowledge may help them to understand their behavior. However, a patient may not want to know they suffer from cancer if there is no available treatment. Whereas another patient may desire the prognosis, even though it is untreatable because they want to make the most of their remaining life. Would you like to know four years in advance that you have a high probability to fall ill in a potential deadly disease? Technology itself is not good or evil, humans are. AI in medicine is a revolution that can either destroy or save the world. It is up to us.

References

1. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. *Nat Methods*. 2018;15(4):233–4. <https://doi.org/10.1038/nmeth.4642>.
2. Ebigo A, Palm C, Probst A, Mendel R, Manzeneder J, Prinz F, de Souza LA, Papa JP, Siersema P, Messmann H. A technical review of artificial intelligence as applied to gastrointestinal endoscopy: clarifying the terminology. *Endosc Int Open*. 2019;07(12):E1616–23. <https://doi.org/10.1055/a-1010-5705>.
3. Fournier A, Fussell D. Stochastic modeling in computer graphics. 1980. <https://doi.org/10.1145/800250.807477>.
4. Gu J, Oelke D. Understanding bias in machine learning. 1st Workshop on Visualization for AI Explainability in 2018 IEEE Vis, 2019..

5. Gunning D. DARPA's explainable artificial intelligence (XAI) program. 2019. <https://doi.org/10.1145/3301275.3308446>.
6. Guo W. Explainable Artificial Intelligence for 6G: improving trust between human and machine. *IEEE Commun Mag.* 2020;58(6):39–45. <https://doi.org/10.1109/mcom.001.2000050>.
7. Kerasidou A. Artificial intelligence and the ongoing need for empathy, compassion and trust in healthcare. *Bull World Health Organ.* 2020;98(4):245–50. <https://doi.org/10.2471/blt.19.237198>.
8. Kumar RSS, O'Brien D, Albert K, Viljoen S, Snover J. Failure modes in machine learning systems. 2019.
9. Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller K-R. Unmasking Clever Hans predictors and assessing what machines really learn. *Nat Commun.* 2019;10(1):1096. <https://doi.org/10.1038/s41467-019-08987-4>.
10. Norkin VI. Method of generalized gradient descent. *Cybern Syst Anal.* 1986;21(4):495–505. <https://doi.org/10.1007/bf01070609>.
11. Reynolds CW. Flocks, herds and schools: a distributed behavioral model. 1987. <https://doi.org/10.1145/37401.37406>.
12. Savage N. How AI is improving cancer diagnostics – Artificial intelligence can spot subtle patterns that can easily be missed by humans. *Nature Outlook*, 25 March 2020. <https://doi.org/10.1038/d41586-020-00847-2>.
13. Schönberger D. Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications. *Int J Law Inf Technol.* 2019;27(2):171–203. <https://doi.org/10.1093/ijlit/eaz004>.
14. Yala A, et al. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology.* 2019;292:60–6. <https://doi.org/10.1148/radiol.2019182716>.
15. Yu AC, Eng J. One algorithm may not fit all: how selection bias affects machine learning performance. *RadioGraphics.* 2020;40(7):1932–7. <https://doi.org/10.1148/rg.2020200040>.



The New Frontiers of AI in Medicine

7

Pritesh Mistry

Contents

Introduction	116
Natural Language Processing	116
Can Future AI Systems Use Natural Language Processing to Improve How Individuals Engage in Their Health?	116
Unlocking Clinical Knowledge to Address Common Patient Concerns Through Clinical Avatars	117
Engineering the Next-Generation Electronic Health Record and Improving Staff Workflows	119
Machine Learning, Deep Learning, and Neural Networks	119
Improved Disease Detection Through Computationally Enabled Diagnostics	119
Entirely New Ways of Diagnosing Disease Using Digital Biomarkers	121
AI-Driven Population-Level Interventions for Entire System Planning and Optimization	121
Computer Vision	122
Using AI to Generate Real-Time Synthetic Images to Improve Surgical Interventions	122
AI-Driven Surgical Success Measures to Create Intelligent System-Wide Resource Planning	122
Robotics	123
Partly Automated Surgery for Rapid Recovery and Safety	123
New Frontiers in AI	124
AI Bias Gives Another Perspective on Driving Improvement	124
Evidence-Based Individualized Treatment Pathways	125
The Future of AI in Medicine	125
References	126

Abstract

This chapter reflects upon the research and innovation at the forefront of artificial intelligence (AI) from hardware to software and their application to draw the potential future

P. Mistry (✉)
The Kings Fund, London, UK

applications of AI that will change how care is delivered irrevocably. Techniques including machine learning, natural language processing, and computer vision will be applied to enable earlier diagnosis, give patient control, and create entirely new categories of diagnostics. AI has the potential to not just digitalize what healthcare currently does but provide uniquely different ways forward that will revolutionize care delivery.

Keywords

Future · Machine learning · Natural language processing · Bias · Digital biomarkers · Automation · Wearables · Literacy · Chatbot · Robotics

Introduction

This chapter is intended to give the reader a glimpse of how AI may alter health and care services in the longer term. The day-to-day changes wrought through AI-powered tools may be dramatic or subtle; either way, the impact is anticipated to be far-reaching.

Three things fuel AI progress – data to inform and train software, algorithms using data to replicate “intelligence” in software, and hardware that runs the algorithms. All three of these areas continue to advance at pace which in turn will accelerate AI applications and their impact. Data is becoming increasingly accessible in healthcare systems, and many healthcare systems are tackling interoperability and transitioning to integrated care providers [1]. This new approach aims to dissolve silos in healthcare systems through the integration of primary, secondary, community, and other health and care services. As integrated care providers mature, the result will be integrated datasets and, alongside consumer devices, richer more comprehensive data pools that can be used to train AI algorithms – improving AI performance. Hardware development is also progressing at pace, and companies are developing bespoke silicon chips specifically designed for AI applications. These chips are being designed for use in every computing device from the phones in our pockets to the tablets and

laptops in our bags, to the computers on our desks and data centers for cloud computing. These AI-optimized silicon chips (called AI accelerators) will continue to push the boundaries of speed and AI complexity across devices. Existing algorithms will get faster, and algorithms that did at one time take too long now become practical. Finally, new AI techniques continue to be developed and refined improving efficiency and functionality. When combined with more specialization it means algorithms become less multi-purpose and more tailored toward a specific application improving efficiency. AI is now being applied to develop the next generation of AI algorithms and techniques [2]; as this continues to develop, new AI techniques are likely to emerge and improve on speed, accuracy, and ability to handle increasingly more complexity.

We can speculate upon how the combination of comprehensive rich datasets with optimized faster hardware and new algorithms with new techniques will unlock AI functionality that will look like magic. Perhaps these will even pave the way to the development of general AI, and the dream of the AI doctor will become a reality. However, this chapter will remain grounded in current research to shine a light on potential developments that are tantalizingly out of reach yet possible in the near future. At the end of the chapter, we will briefly cover how AI may unlock totally new ways of working.

Natural Language Processing

Can Future AI Systems Use Natural Language Processing to Improve How Individuals Engage in Their Health?

Healthcare systems are moving toward a more person-centered model of care provision with health literacy, patient activation measure, and shared decision-making becoming more important as a way of categorizing the capability, willingness, and approach of co-deciding care alongside the patient.

A patient with good literacy skills is an individual empowered to be in control of their health, care, and wellbeing. Literacy is essential for being

able to understand the situation and be confident in making decisions. Literacy underpins accessing self-help, understanding consultations, engaging in remote consolations, researching and digesting background information from trusted sources, reading and understanding test results, being able to understand medical records, interpreting medical guidance and medicine side effects, and uncountable more aspects of health and care.

The literacy capability challenge is twofold: i) patients' literacy levels need to be adequate enough that they can convey information and comprehend what they are reading (e.g., in records, test results, or patient information leaflets) or told (e.g., in consultations or conversations); and ii) clinicians need to be able to gauge a patient's literacy level and match this in all forms of communication. AI has the potential to alter radically shift information exchange improving understanding and patient engagement.

Natural language processing (NLP) essentially takes speech (or text) that's constructed with natural sentence structure through normal dialogue and converts this speech into text (or text into speech). Currently NLP is capable of taking a range of natural language sentence structures and words to confer a generalized meaning irrespective of how the original sentence is structured or the variety of the words used, e.g., NLP can assume the equivalence of lots of pain, painful, really hurts, and acute pain.

Current linguistic research is exploring how language and meaning change across different academic disciplines. NLP tools are improving the understanding of language and communication, research findings show that subjects as different as physics and history have similarly complex features [3]. This research is improving our knowledge of how language can be restructured to be interpretable at varying degrees of complexity. There is complementary work bringing together communication scientists, computational linguistics researchers, and health service researchers and staff; it shows promising indications of where language complexity and NLP may lead in the future. The research demonstrates NLP tools can analyze and interpret patient communications to gauge a person's health literacy level with good accuracy (>80%) [4]. The tools can

also interpret clinician communication in real time and advise if the health literacy level is incompatible with the patients' capabilities [5].

As this strand of NLP research and tools progress in the future, it will be possible to have a more nuanced interpretation of the natural language sentence structures within health and care. This will create NLP software tools that are able to span different levels of literacy and sentence structure sophistication (see Fig. 1). What this means in reality is that clinical information could be translated into audience-appropriate text, speech, or visual media depending upon the literacy level of the individual. This use of AI will empower individuals to access understandable health and care information in their records, medication guidance, and trusted sources without requiring additional staff resources. This development of NLP will be the bridge joining an individual's capabilities with specialized information.

Unlocking Clinical Knowledge to Address Common Patient Concerns Through Clinical Avatars

A positive diagnostic test for a life-altering or life-threatening condition commonly leaves patients feeling disorientated. They need time to process the news, consider the implications, and come to terms with the diagnosis. Directly upon receiving the news, the patient will have immediate questions, but more often, they or their family have many more questions which spring to mind hours, days or even weeks after the consultation. Patients either need to wait for another contact with the clinician or attempt to navigate the information available through the Internet or patient information leaflets. Waiting for another appointment can increase stress and anxiety, which is detrimental to their mental health. Furthermore, an appointment to answer routine questions might not be the most effective use of clinical time during an appointment. Instead, the clinical time could be better used if the patient has their questions answered before their next appointment enabling a more therapeutic and informed consultation. Navigating patient information leaflets or information on the Internet is an option but

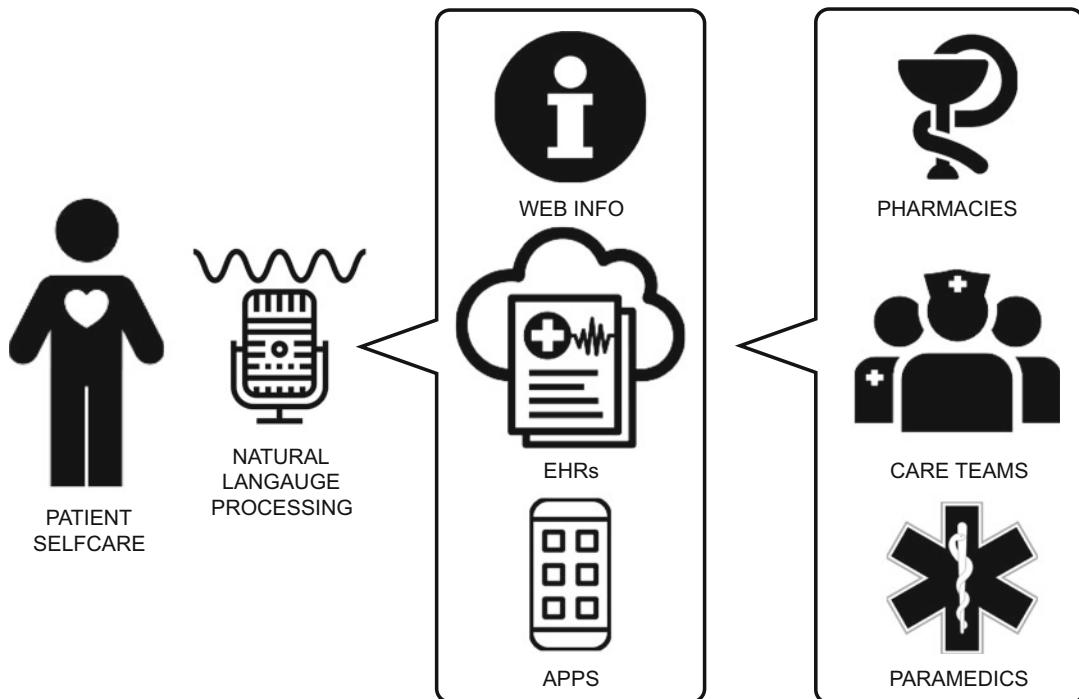


Fig. 1 Natural Language Processing (NLP) will enable clinical information to be translated to meet the readers capabilities and needs (Illustration: Dr Niklas Lidströmer)

requires patients to have a good enough level of literacy (as discussed above) and be able to find trusted sources of information. The literacy requirement excludes patients, while incorrect information has the potential to cause more harm.

AI in the future has the potential to change this to improve patients' knowledge and how clinical time is used. NLP could be applied to create clinical chatbots for routine information [6]. Chatbots are now very common, but the quality is highly variable, with many of us having very good and very frustrating experiences.

In the future, the technology will be mature enough for specialists, or hospitals, to be able to develop chatbots that act as clinical avatars available 24 h 7 days a week to answer patients' questions on their recent diagnosis or ongoing treatment [7]. This has the potential to help patients navigate their change in health and have timely trusted information while saving clinical time. Such avatars will be able to provide factually correct responses to a patient's queries when a patient needs it. Orientating these tools through

healthcare organizations will enable the combined clinical knowledge and expertise of multiple specialists to be used to train the avatar sharing the burden to create and maintain the avatars and also increasing the scope of the information available. A clinical chatbot would need additional oversight and training to mitigate against medical risk and harm. However, the benefits include saving clinical time and improving accessible trusted knowledge for patients for routine questions in common clinical areas.

There is increasingly more research focus on therapeutic empathy and the potential for machines and software to have perceived empathy [8]. Empathy is an essential skill for therapeutic consultations. While public sentiment is generally quite skeptical about machines exhibiting empathy, research suggests that people naturally project social rules and expectations on machines [9] which contribute to perceived empathy. When engaging with chatbots, people prefer health advice chatbots that exhibit sympathy and empathy [10]. Longitudinal studies demonstrate people build relationships with

chatbots; however, the repetitive and predictive nature of current chatbots means relationships are not long-lasting. As chatbot functionality continues to mature, clinical avatars will take on a conversational therapeutic function for some clinical situations and conditions for some individuals. It is important to note this is very different to an AI doctor, clinical avatars have the potential to provide short-term support through information and a form of treatment through therapeutic conversations, but they are likely to continue to be capable of functioning in narrow application areas and a limited number of situations.

Engineering the Next-Generation Electronic Health Record and Improving Staff Workflows

Electronic health records (EHRs) are longitudinal records of a person's health and their interaction with the health system. They have been created and implemented in different ways in different healthcare systems for slightly different purposes – e.g., care provision, reimbursement, etc. Capturing the information takes the time and attention of highly qualified clinicians; studies show out of an 11.4-h day, approximately 6 h of clinician time is used for data entry [11]. It is believed that EHRs are time-consuming and disrupt workflows to such an extent that they are contributing to increased pressure on clinicians resulting in burnout [12].

EHRs are essential, they put crucial information at the clinicians' fingertips to improve the quality of care and reduce medical errors, but they also contain bias and variation in detail. Different clinicians take different approaches to recording patients' information. Some clinicians record information verbatim; others record interpretations [13]. Similarly, consultants dictate or record summaries for support staff who create the records and/or relating correspondences (e.g., letters) creating variation. Furthermore, EHRs hold multiple different types of structured and unstructured information, from test results to consultation notes to specialist clinic letters and images. All these in combination mean that there's a

significant degree of variation in the information quality and a range of information types.

Natural language processing (NLP) is being developed to reduce some of this administrative burden on healthcare workers [14]. The speech-to-text capability of NLP has the potential to be used to record and categorize a patient's electronic health record without the direct involvement of the clinician. A clinician can dedicate their time in consultation instead of multi-tasking. This reduces cognitive load decreasing burden and improving the quality of consultation time. It is anticipated this will improve the quality of care and save time for the healthcare staff. The information captured and coded would be more consistent across staff and provider organizations creating the opportunity for improved research and analysis for population health interventions, for example. It will also be possible to use the analysis of vocal patterns as a new type of indicator of clinical condition.

Along with the time savings and the improvement in consultations, there may be unanticipated consequences. The process of creating EHR entries is believed to help clinical staff to synthesize and assimilate the information disclosed by the patient [15]. Using NLP for EHR entry processes could impact how staff develop clinical judgment and clinical knowledge; changes to medical training will be required to circumvent any detrimental effect on clinical judgment.

Machine Learning, Deep Learning, and Neural Networks

Improved Disease Detection Through Computationally Enabled Diagnostics

Biomarkers are the rockbed of modern medicine, and biomarkers that are well characterized have repeatedly demonstrated that they correctly predict clinical outcomes across a range of treatments and populations [16]. This means an individual's physiological state can be monitored or managed with confidence using one or more biological markers.

A biomarker was defined in 1998 by the National Institutes of Health Biomarkers Definitions Working Group [17] as “a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention.”

Generally biomarkers are measured and monitored with diagnostic medical devices. These are devices that accurately measure one or more biomarker, and the accuracy is verified through evidence that is generated using appropriate methodologies (such as a clinical trial). The devices are checked that they are acceptably low risk through adherence to medical device regulations. Non-medical devices such as wearables do not have this rigor in evidence and regulation, but there is potential for non-medical devices to be used to change how diseases are detected and monitored.

Wearables are devices that are not designed to be medical devices but consumer wellness or fitness devices. The sensors found on wearable devices often measure medical biomarkers such as heart rate, breathing rate, ECG, blood oxygenation, etc. But wearables do not undergo the level of scrutiny, evidence generation, or regulatory approval that is expected of medical devices. Wearable devices have a number of unique aspects compared to medical devices which could significantly alter how diseases are detected, treated, and managed. Wearables generate data in much higher quantities than medical devices; this is because they have high sampling frequencies and are always on the person. They also potentially have higher levels of variability and errors – the accepted variation between devices is greater for fitness purposes than health applications, manufacturing tolerance can be larger, devices do not have mandated regular maintenance and may be jostled when worn [18]. So, while there are huge amounts of data, there are also significant noise, irrelevant data, and poor-quality data.

AI is the ideal tool to create meaning from the vast quantities of wearable data. With AI, data analysis can be automated by rejecting abnormalities, flagging non-compliance, and reducing variance to provide data that is interpretable and

then apply analysis to derive clinically meaningful insights. This field which can be considered as computationally enhanced diagnostics is in many ways analogous to computation photography (see Box 1).

Box 1 Computational Photography

AI has unlocked a revolution in smartphone photography – tiny cameras, powerful silicon chips, and AI in combination enable images to be created which are far better than raw camera images. AI techniques process the raw camera image, and their application enhances and improves the quality and features of the images captured.

AI-powered software will effectively interface between the wearables that measure biological markers and time-constrained healthcare staff who seek actionable insights with which to support their patients. These AI software tools at the most basic level filter the jostles and noise to leave the usable data. As AI advances, it will create personal baselines for everyone individualizing the detected changes and create inferences by adding other data sources (e.g., location, pollen, pollution, temperature, etc.) to link triggers with symptoms. These in combination will complement existing medical diagnostics to improve the diagnosis, treatment, and management of clinical conditions. Information will be more relevant to the individual and actionable to address deterioration or symptoms.

This AI software could dramatically change healthcare and medical knowledge. Access can take a much more personalized approach based on data related to physiological or psychological changes instead of the current structure of routine appointments. Wearables will give frequent data sampling across broad demographics, and this will change how disease is measured. Firstly, the quantity of data gives more insight on range of variation in clinical conditions. Secondly, the ability to measure biomarkers in real-world circumstances. The ramifications of this data and understanding is a change in the very nature of how disease is defined if high degrees of

individual variability are found. For example, it will raise questions on what constitutes as the boundary between ill-health and good health.

Entirely New Ways of Diagnosing Disease Using Digital Biomarkers

Digital devices (phones, tablets, computers, etc.) are very different to most wearables; our digital devices contain sensors that measure what we do but do not measure the generally accepted clinical biomarkers. Instead, digital devices measure (or can measure) more mundane things such as how we are typing or communicating. It's possible that this more mundane activity can give insights into our wellbeing and our health conditions.

How we engage and use digital devices has the potential to create new biomarkers – digital biomarkers (see Fig. 2) – that give indications on our health. For example, the gait of the way we walk changes with certain diseases [19] and can be detected using machine learning, the way we tap

out a text message can change with the onset of tremors [20], how we cough can indicate the type of respiratory illness [21], and our phone usage behavior can indicate our mental health [22].

Our digital behaviors indicate our physiological and psychological wellbeing; this can be utilized to create new types of biomarkers for the early detection of disease and ongoing disease monitoring. Similar to wearables, the data generated from these digital biomarkers are variable and noisy, and AI is essential to identify the usable data and link the various data types to create clinically meaningful digital biomarkers. Early indications are that digital biomarkers are unique to an individual [22] which ordinarily would pose a problem to train machine learning algorithms, but the quantity of data each person generates should avoid data limitation problems. With the right machine learning algorithms, it will be possible to track how an individual engages with their devices, examining how we scroll, swipe, tap, etc.; creating a personal baseline for each type of engagement action; and using meaningful divergence from the baseline to indicate deterioration in health and wellbeing.

AI is well suited for such a new approach for diagnosing illness, and it requires tracking and analyzing thousands of data sources daily for each individual and across multiple devices and with a huge number of variables. Approaches such as machine learning, deep learning, and neural networks are ideal to analyze vast datasets so that we can identify patterns and establish new patterns or deviations from existing patterns. It is these new patterns and deviation from existing patterns that would firstly generate evidence that these digital biomarkers give indications of clinical conditions and help provide timely indication of a meaningful clinical change in the individual.

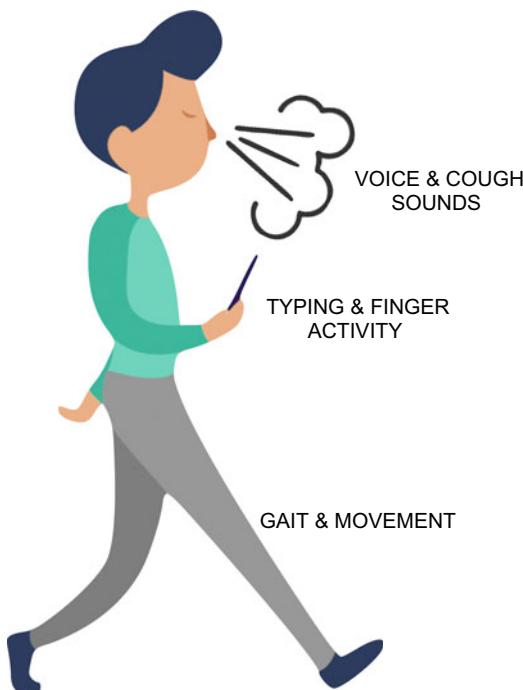


Fig. 2 Our use of digital devices has the potential to create a new generation of biomarkers (Illustration: Dr Niklas Lidströmer)

AI-Driven Population-Level Interventions for Entire System Planning and Optimization

The quantity of data generated from one wearable is vast, and the above section demonstrates the need for AI to make sense of it all and create

actionable insight on a person's health. The quantity of data collected across a geographical region (such as that used for population health management) is orders of magnitude larger. This huge data quantity across a regional population has the potential to power population health interventions and public health initiatives complementing information from healthcare records. However, the challenge of using such a large volume of data is also considerable.

Machine learning algorithms will be able to support the data analysis by identifying patterns through linked data. For example, AI software can be applied to electronic health records and consumer devices to create insight on the health trends for local populations and within particular clinical conditions. This will help to target local-level interventions for particular clinical conditions in the first instance it could inform a need to improve particular care pathways. Fusing various sources of health and environmental information (e.g., air pollution, pollen, noise pollution) has the potential to help forecast service capacity changes, e.g., medication usage, mental health needs, social prescribing needs, etc.

Computer Vision

Using AI to Generate Real-Time Synthetic Images to Improve Surgical Interventions

Currently, AI software applied to medical imaging is able to improve the quality and speed of specific parts of the diagnosis process. AI within hospitals has demonstrated the potential to improve the speed of segmentation for clinical diagnosis and quantify changes in physiology (e.g., tumors, retinopathy, dermatology, etc.) [23]. AI is unlocking specialist tools and making them available in the hands of the public; it is now possible to detect diabetic retinopathy [24] and identify melanoma [25] using AI-powered software and a smartphone.

AI techniques have the potential to improve medical images through feature enhancement and interpolating features to improve the resolution of small structures beyond the capability of

the hardware. Computational techniques will also make it possible to artificially increase the contrast of low-dose live x-ray images which would reduce exposure to ionizing radiation while improving visibility for the surgical team.

As AI-powered imaging systems improve, the benefits will compound. Currently, during surgical planning, images are taken using several modalities. Commonly in, for example, cardiac surgery, live x-ray is used for tracking surgical instruments, and past images from other modalities are overlaid to aid the intervention. In the near future, software will be able to automatically combine images from different modalities to create fused images that merge the advantages of each modality into a composite image [26]. If bias and error generation issues can be solved, it will pave the way for more sophisticated AI tools trained on pre-operative images and video and combined with live x-ray during interventions. Such tools will be able to reconstruct medical images and video, and so they will simulate live videos composed of information from several modalities and display this to surgical teams during interventions. Such AI imaging fusion will improve capture speed, quality, and accuracy of the images enabling lower functionality equipment to be used for higher sophistication purposes. It will also enable lower dosage of radioisotopes and x-ray radiation improving safety and reducing costs.

AI-Driven Surgical Success Measures to Create Intelligent System-Wide Resource Planning

In the near future, machine learning applied to electronic health records will identify correlations between surgical intervention and recovery times predicting the requisite post-operative rehabilitation support. This could be used at surgical planning stages to forecast the future resources that will be required such as community support and rehabilitation required.

The accuracy of such a tool would vary considerably with a dependency upon multiple variables not just the intervention type limiting the utility. The use of computer vision will further

enhance these predictive models so they can use medical imaging from surgical interventions to provide additional data to train models. The use of image features from surgery combined with electronic health records data will give an improved forecast of post-operative needs of the patient. The combination of retrospective image data with information on procedure outcomes will enable forecasting of resources required for successful recuperation; this would aid budgetary forecasting and enable improvements to efficacy. These models could then be iterated upon, incorporating rehab and fitness data to improve the projections in near real time.

Robotics

Partly Automated Surgery for Rapid Recovery and Safety

Robots for robotic surgery have been available on the market for over a decade [27]. The robots however are not automated, that is, they always

require a human operator – usually in the same room or an adjoining room. This limits many of the potential benefits of robotic surgery, and when combined with hardware limitations, it means that they have only been utilized in niche areas since becoming widely available.

More recently, with the aid of high-fidelity Internet connectivity, tele-robotic surgery has started to emerge [28]. In this case, the human operator can theoretically be anywhere in the world and can remotely connect to the robotic equipment to remotely operate on a patient.

The advent of tele-robotic surgery has unlocked access to additional information and data alongside the previously accessible medical image data (see Fig. 3). It will soon be possible to compile these data to automate one or more steps of an intervention, but it might not be possible or advisable to automate entire procedures. For example, through computer vision techniques, it will be possible to use AI to plan the incision step of an intervention automating a robot to carry out this part of the process potentially reducing scarring and reducing recovery times.

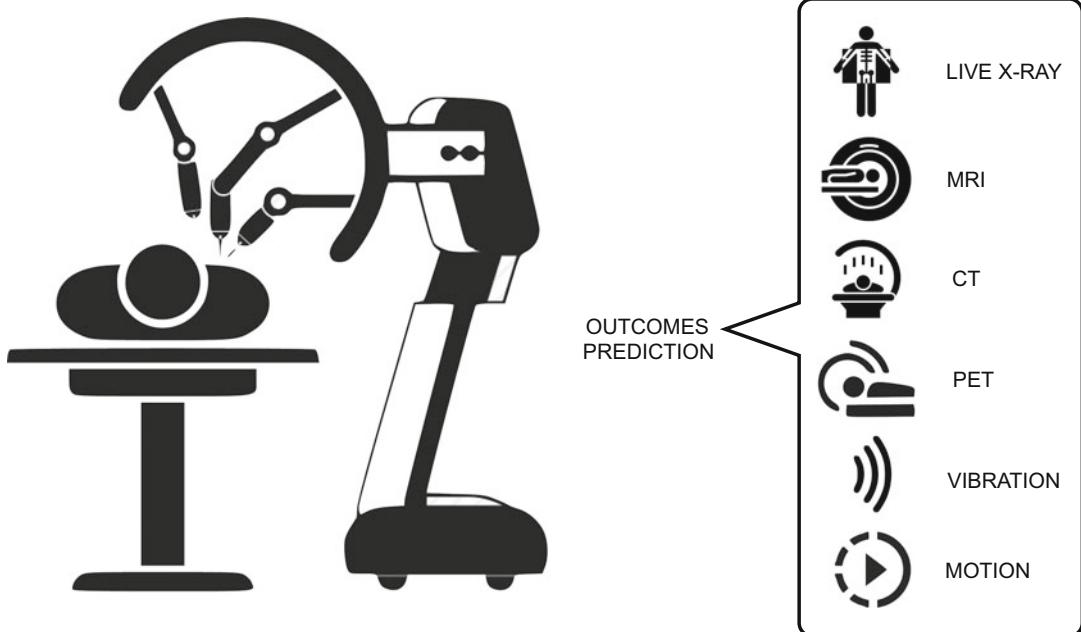


Fig. 3 Remote robotic surgery utilising combining stored and live images with predictors and tracking to estimate outcomes and rehab needs (Illustration: Dr Niklas Lidströmer)

In the long term, it might be possible to automate entire interventions; in the short to medium term, steps of highly routine interventions will become automated and use robotics; but in the medium term, it's unlikely that entire operations will be automated. In areas such as congenital heart disease, the variability in condition and intervention approach creates a high risk of unforeseen emergencies this means we will need humans to be involved and engaged in the surgical interventions. Semi-automation reduces human awareness and perception in the moment, while perceived high-fidelity systems can result in less vigilance from human operators both of which can increase harm (see Box 2).

Box 2 Self-Driving Cars

The early 2000s onward has seen an explosion in the development of autonomous vehicles (or self-driving cars). The combination of sensors, computing power, and software has resulted huge leaps in capability progressing toward automating vehicles, but they are still only semi-automated. These semi-autonomous vehicles have now travelled millions of kilometers on public roads improving the algorithms and capabilities with each journey. Along with the training, there have been accidents, sometimes despite a person sitting in the driver seat [29]. The reason for this is a lack of engagement in the action of driving which reduces attentiveness requiring longer to respond to a rapidly changing situation. This can also give insight into the risks of semi-automated surgery.

In a driving situation that rapidly changes, it's been documented that a person not actively controlling the vehicle has difficulty re-establishing context (referred to as "rapid onboarding" by psychologists); it's essential the person take control of the vehicle as the automated system is unable to handle the situation. However, it's likely the human driver takes time to re-establish situation awareness critical time that can make the

difference to whether there is an incident. Research shows the onboarding times grow longer when high levels of automation are combined with complex situations [30]. Furthermore, there is growing evidence that prolonged use of automation deteriorates human skills, with cognitive skills deteriorating quickly, but manual skills remain. However, the cognitive skills are essential to use our manual skills effectively [31].

AI-augmented surgeons carrying out surgery using robotics could improve patient safety and reduce adverse incidents. The use of machine learning with computer vision could aid the development of AI systems that are able to analyze intervention images in real time and reduce operator hand tremors. The same images with AI interpretation could establish safety based AI algorithms which guard against commonly avoidable surgical errors [32].

New Frontiers in AI

AI Bias Gives Another Perspective on Driving Improvement

AI has the potential to be a technology that knits together other technologies and unlocks insight that has until now been inaccessible. With all such technologies, their immediate application replicates the status quo – the current way of doing things. However, in successive generations of AI, the unique capabilities will get used more and will completely change how we approach and deliver healthcare.

Healthcare Service Redesign

Often AI is considered to be a tool that solves a problem, that is, we feed data in, and it gives us an answer. However, data has historically been used to identify areas that warrant a closer look. AI can be used for this approach to power a learning and evolving healthcare system that uses AI to suggest quality improvement projects to reduce systemic inequalities.

AI is a data-driven software tool, the algorithms and software are written by people, and this can include biases. Additionally, the data

collected to train systems may replicate these biases or introduce new ones. Using such data-driven software risks perpetuating biases, hard coding them into complex software applications, and embedding and exacerbating inequalities [33]. It is absolutely essential that the AI software are evidence based and regulated to ensure acceptable levels of risk. Equally important, if not more so, is to ensure the healthcare system has the knowledge, processes, skills, and tools (including datasets) to audit and interrogate AI software for bias. Firstly, this would ensure biased AI is identified prior to being implemented so the exact use case of the software can be defined; otherwise, there is a risk of widening inequalities or even harm [34]. Secondly, with this knowledge, it may be possible that AI can help improve healthcare services in novel ways. By their nature of being biased, AI can provide insight into where there are inequities in service provision.

AI applied to local population-level data could be used as a healthcare service redesign and improvement tool to help address inequalities within particular clinical conditions by helping to monitor pathway effectiveness for particular cohorts or surface a pattern of need that is currently unidentified.

This information can help to identify everyday services and pathways which can be improved for specific populations to continue to develop an equitable healthcare system.

Evidence-Based Individualized Treatment Pathways

There's potential for AI to power a next generation of clinical decision-making, one that supports healthcare staff to use all the interventional tools available to support patients.

For a long time, treatment approaches have been quite simple with a range of drugs, therapies, or surgeries. With progress comes a proliferation of medications, therapies, and surgical interventions all with different efficacies, risks of side effects, and harm. The progress in genomics has compounded the complexities in selecting the right medication for an individual. Drugs have

been combined with devices for common surgical interventions like angioplasty. There are a growing interest and evidence base around the use of social prescribing – social engagement to treat non-communicable diseases. Digital therapeutics (apps as an intervention and treatment approach) is another area with growing interest and evidence. It's not possible for a healthcare worker to know about all of these and how they can be fit together to improve a patient's health. The default is to rely on familiar treatments which might not be the best suited for that particular individual.

AI is a well-suited tool to help make sense of all the potential treatment and intervention choices to create options (see Fig. 4). AI can identify all the treatments with evidence that match the individuals' demographics giving them the best possible outcomes and align to their preferred treatment methodology. AI can filter, adjust, and overlay these multiple approaches to suit the patient's preferences and goals. When combined with the clinical risk and need this presents a portfolio of options of essentially personalized pathways that can be selected from by the healthcare staff with the patient, to find an approach which would best suit the patient. Using the most effective combination of medicines, surgery, therapies, digital tools, and community activities available not just the tools most familiar to the consulting clinician.

The Future of AI in Medicine

This chapter outlines how AI has the potential to substantially change and improve the delivery of healthcare. As covered in the beginning of the chapter, AI can improve how we currently think of healthcare system and delivery. The two theoretical scenarios provided at the end of the chapter demonstrate that AI has huge potential to change how we design healthcare services in ways that have never been seen before. In the short term, we will see AI tools that replicate current care provision identifying features in images, symptom checking, creating electronic health record entries, etc. In the medium to long term, we will see completely new ways of creating patient pathways, monitoring disease, and improving the

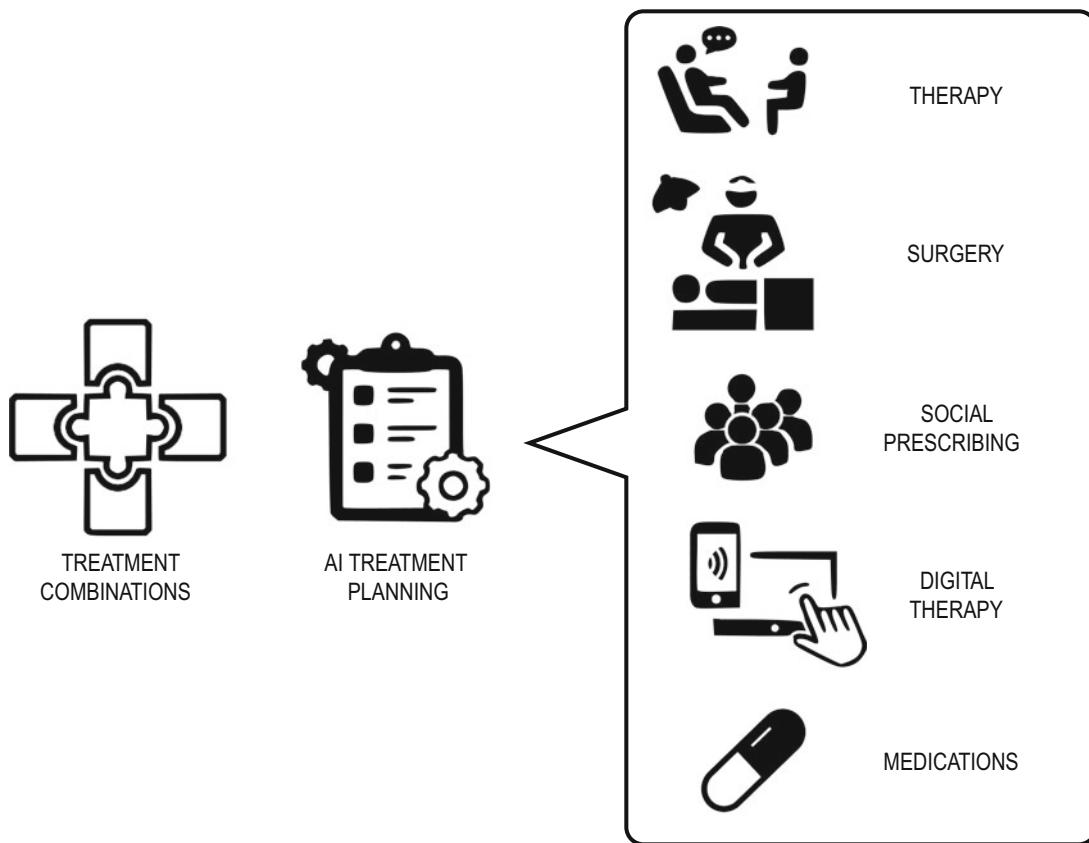


Fig. 4 AI which combines multiple evidence and service information sources to provide options for combinations of treatment approaches (Illustration: Dr Niklas Lidströmer)

healthcare service. Much of the AI development discussed in this chapter is at the forefront of what is possible; it challenges processes and standard guidelines. Similar to the Wild West, and with all new frontiers, the rules need to be adapted and applied to minimize risk, remove bias, and ensure evidence, and high quality is at the core of AI applied to medicine.

References

1. Delnoij D, Klazinga N, Glasgow IK. Integrated care in an international perspective. *Int J Integr Care*. 2002;2:e04. <https://doi.org/10.5334/ijic.62>.
2. Real E, Liang C, So D, Le Q. AutoMLZero: evolving machine learning algorithms from Scratch, arXiv, Mar 2020, [online] arXiv:2003.03384v2. <https://arxiv.org/abs/2003.03384>
3. Green C. A multilevel description of textbook linguistic complexity across disciplines: leveraging NLP to support disciplinary literacy. *Linguist Educ*. 2019;53:100748. <https://doi.org/10.1016/j.linged.2019.100748>.
4. Crossley S, Liu J, Karter A, McNamara D, Schillinger D. Developing and testing automatic models of patient communicative health literacy using linguistic features: findings from the ECLIPPSE study. *Health Commun*. 2020;36:1018–28. <https://doi.org/10.1080/10410236.2020.1731781>.
5. Schillinger D, McNamara D, Crossley S, Lyles C, Moffet HH, Sarkar U, et al. The next frontier in communication and the ECLIPPSE study: bridging the linguistic divide in secure messaging. *J Diabetes Res*. 2017;2017:1348242. <https://doi.org/10.1155/2017/1348242>.
6. Hawthorne K, Connor J, Taubert M, Murphy D. Symposium B6: will artificial intelligence support new approaches to health which will empower patients within the next five years? 2019.
7. Amato F, Marrone S, Moscato V, Piantadosi G, Picariello A, Sansone C. Chatbots meet eHealth: automatizing healthcare. *CEUR Workshop Proceedings*. 2017. <http://ceur-ws.org/Vol-1982/paper6.pdf>. Accessed 2021-01-35.

8. Howick J, Bizzari V, Dambha-Miller H. Therapeutic empathy: what it is and what it isn't. *J R Soc Med.* 2018;111(7):233–6. <https://doi.org/10.1177/0141076818781403>.
9. Nass C, Moon Y. Machines and mindlessness: social responses to computers. *J Soc Issues.* 2000;56:81–103. <https://doi.org/10.1111/0022-4537.00153>.
10. Liu B, Sundar SS. Should machines express sympathy and empathy? Experiments with a health advice chatbot. *Cyberpsychol Behav Soc Network.* 2018;21(10):625–36. <https://doi.org/10.1089/cyber.2018.0110>.
11. Arndt BG, Beasley JW, Watkinson MD, Temte JL, Tuan WJ, Sinsky CA, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. *Ann Fam Med.* 2017;15(5):419–26. <https://doi.org/10.1370/afm.2121>.
12. Noseworthy J, Madara J, Cosgrove D, Edgeworth M, Ellison E, Krevans S, et al. Physician burnout is a public health crisis: a message to our fellow health care CEOs. 2017. <https://www.healthaffairs.org/do/10.1377/hblog20170328.059397/full/>, Accessed 27 Jan 2021.
13. Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible sources of bias in primary care electronic health record data use and reuse. *J Med Internet Res.* 2018;20(5):e185. <https://doi.org/10.2196/jmir.9134>. <https://www.jmir.org/2018/5/e185>
14. Langston J. Microsoft and Nuance join forces in quest to help doctors turn their focus back to patients. 2019. <https://blogs.microsoft.com/ai/nuance-exam-room-of-the-future/>. Accessed 27 Jan 2021.
15. Willis M, Duckworth P, Coulter A, Meyer ET, Osborne M. The future of health care: protocol for measuring the potential of task automation grounded in the National Health Service Primary Care System. *JMIR Res Protoc.* 2019;8(4):e11232. <https://doi.org/10.2196/11232>.
16. Strimbu K, Tavel JA. What are biomarkers? *Curr Opin HIV AIDS.* 2010;5(6):463–6. <https://doi.org/10.1097/COH.0b013e32833ed177>.
17. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther.* 2001;69(3):89–95. <https://doi.org/10.1067/mcp.2001.113989>.
18. de Arriba-Pérez F, Caeiro-Rodríguez M, Santos-Gago JM. Collection and processing of data from wrist wearable devices in heterogeneous and multiple-user scenarios. *Sensors (Basel).* 2016;16(9):1538. Published 2016 Sept 21. <https://doi.org/10.3390/s16091538>
19. Rehman RZU, Del Din S, Guan Y, et al. Selecting clinically relevant gait characteristics for classification of early Parkinson's disease: a comprehensive machine learning approach. *Sci Rep.* 2019;9:17269. <https://doi.org/10.1038/s41598-019-53656-7>.
20. Parra V, Figueras G, Huerta M, Marzinotto A, Gonzalez R, Alvizu R. A smartphone application for Parkinson Tremor detection. Conference workshop IEEE Lantincom, 2013.
21. Porter P, Abeyratne U, Swarnkar V, et al. A prospective multicentre study testing the diagnostic accuracy of an automated cough sound centred analytic system for the identification of common respiratory disorders in children. *Respir Res.* 2019;20:81. <https://doi.org/10.1186/s12931-019-1046-6>.
22. Place S, Blanch-Hartigan D, Rubin C, Gorrostieta C, Mead C, Kane J, et al. Behavioral indicators on a Mobile sensing platform predict clinically validated psychiatric symptoms of mood and anxiety disorders. *J Med Internet Res.* 2017;19(3):e75. <https://doi.org/10.2196/jmir.6678>. <https://www.jmir.org/2017/3/e75>
23. Dey D, Slomka P, Leeson P, Comaniciu D, Shrestha S, Sengupta P, et al. Artificial intelligence in cardiovascular imaging: JACC state-of-the-art review. *J Am Coll Cardiol.* 2019;73(11):1317–35. <https://doi.org/10.1016/j.jacc.2018.12.054>.
24. Tan CH, Quah W, Tan CSH, et al. Use of smartphones for detecting diabetic retinopathy: a protocol for a scoping review of diagnostic test accuracy studies. *BMJ Open.* 2019;9:e028811. <https://doi.org/10.1136/bmjopen-2018-028811>.
25. Phillips M, Marsden H, Jaffe W, et al. Assessment of accuracy of an artificial intelligence algorithm to detect melanoma in images of skin lesions. *JAMA Netw Open.* 2019;2(10):e1913436. <https://doi.org/10.1001/jamanetworkopen.2019.13436>.
26. Zaharchuk G. Next generation research applications for hybrid PET/MR and PET/CT imaging using deep learning. *Eur J Nucl Med Mol Imaging.* 2019;46:2700–7. <https://doi.org/10.1007/s00259-019-04374-9>.
27. Lanfranco AR, Castellanos AE, Desai JP, Meyers WC. Robotic surgery: a current perspective. *Ann Surg.* 2004;239(1):14–21. <https://doi.org/10.1097/01.sla.0000103020.19595.7d>.
28. Evans CR, Medina MG, Dwyer AM. Telemedicine and telerobotics: from science fiction to reality. *Updat Surg.* 2018;70:357–62. <https://doi.org/10.1007/s13304-018-0574-9>.
29. Casner S, Hutchins E, Norman D. The challenges of partially automated driving. *Commun ACM.* 2016;59(5):70–7.
30. Gold C, Dambock D, Lorenz L, Bengler K. Take over! How long does it take to get the driver back into the loop? In: Proceedings of the human factors and ergonomics society annual meeting (San Diego, CA, Sept 30–Oct 4). Human Factors and Ergonomics Society, Santa Monica, 2013, 19381942.
31. Casner SM, Geven RW, Recker MP, Schooler JW. The retention of manual flying skills in the automated cockpit. *Hum Factors.* 2014;56(8):15061516.
32. Sarker S, Vincent C. Errors in surgery. *Int J Surg.* 2005;3(1):75–81. <https://doi.org/10.1016/j.ijsu.2005.04.003>.
33. Kaushal A, Altman R, Langlotz C. Health care AI systems are biased. 2020. <https://www.scientificamerican.com/article/health-care-ai-systems-are-biased/>. Accessed 27 Jan 2021.
34. Kaushal A, Altman R, Langlotz C. Geographic distribution of US cohorts used to train deep learning algorithms. *JAMA.* 2020;324(12):1212–3. <https://doi.org/10.1001/jama.2020.12067>.



Social and Legal Considerations for Artificial Intelligence in Medicine

8

Matjaž Perc and Janja Hojnik

Contents

Introduction	130
Social Challenges	131
Juristic Challenges	133
Tort Law	134
Conclusions and Guidelines	136
References	137

Abstract

Artificial intelligence is becoming seamlessly integrated into our everyday lives, augmenting our knowledge and capabilities in driving, avoiding traffic, finding friends, choosing the right movie, or finding the perfect song, and, perhaps most importantly, it is entering into

healthcare and medical diagnostics with large brave strides. As this twenty-first century “man meets machine” reality is unfolding, several social and juristic challenges emerge for which we are in general poorly prepared. We here review social dilemmas where individual interests are at odds with the interests of others, and where artificial intelligence might have a particularly hard time making the right decision. Examples thereof are the well-known social dilemmas of autonomous vehicles and vaccination. We also review juristic challenges, with a focus on torts and product liability that are due to artificial intelligence, resulting in the claimant suffering a loss or harm. Here the challenge is to determine who is legally liable, and to what extent. We conclude with an outlook and with a short set of guidelines for constructively mitigating described challenges, with a focus on artificial intelligence in medicine.

Supplementary Information: The online version of this chapter (https://doi.org/10.1007/978-3-030-64573-1_266) contains supplementary material, which is available to authorized users.

M. Perc
Faculty of Natural Sciences and Mathematics, University of Maribor, Maribor, Slovenia

China Medical University Hospital, China Medical University, Taichung, Taiwan

Complexity Science Hub Vienna, Vienna, Austria

J. Hojnik (✉)
Faculty of Law, University of Maribor, Maribor, Slovenia
e-mail: janja.hojnik@um.si

Keywords

Social dilemma · Vaccination · Cooperation · Reward · Punishment · Agent-based model · Tort law · Legal liability

Introduction

A broad body of literature anticipates that in the years to come, intelligent objects will overtake more and more jobs that people have traditionally performed, from driving, diagnosing diseases, providing translation services, and drilling for oil to even milking cows, to name just some examples [1]. In 1999, a British visionary Kevin Ashton coined the term “Internet of Things” (IoT) to describe a general network of things linked together and communicating with each other as computers do today on the Internet [2]. The connection of objects to the Internet makes it possible to access remote sensor data and to control the physical world from a distance [3]. Data communication tools are changing “tagged things” into “smart objects” with sensor data supporting a wireless communication link to the Internet [4, 5]. This means that the manufacturer can make fewer visits, reducing costs and producing less disruption and higher satisfaction for the customer [6]. Remote diagnostics, where complex manufactured products are monitored via sensors, may not, however, only be important for repairing industrial machines but also for human health, such as remote control of pacemakers [7]. The widespread use of Wi-Fi and 4G enables the communication with smart objects without the need of a physical connection, such as to control customers’ home heating and boiler from their mobile or laptop. Mobile smart objects can move around and GPS makes it possible to identify their location [3]. This technology facilitates the development of so-called connected or automated cars that enable the driver automatic notification of crashes and speeding, as well as voice commands, parking applications, engine controls, and car diagnosis. It is foreseen that trucks will soon no longer need drivers, as computers will drive them, without the need for rest or

sleep. Moreover, each Philips or Samsung TV comes nowadays with an application called “Smart TV,” which consolidates video on demand function, the Internet access, as well as social media applications [8]. Objects are thus becoming increasingly smart and consequently autonomous. There are many implications of this in the field of medical law, in particular in relation to the mobile health apps, e.g., a smart phone that is acting as a thermometer or as a blood pressure monitor, applications that track events, retrieve medical content, or allow patient-doctor communication.

However, autonomous objects will also cause accidents, invade private space, fail surgeries and fail to diagnose cancer, and even engage in war crimes [9]. As autonomous objects will become more and more commonplace on streets, on the skies, in households, and in medicine, their social and legal status will only grow in importance. Considering that autonomous objects are not a matter of “if” but rather of “when” such technology will be introduced, the regulatory dimension might be decisive in this respect, as is the prior identification of socially challenging situations where not only the user but also others may be adversely affected. If the activity of autonomous objects, and more generally of artificial intelligence, is not properly regulated, it will not be broadly accepted as a more efficient and safe alternative to human controlled objects or human decision-making. However, the autonomy we give to machines may render many established legal doctrines obsolete, and more importantly, affect what we judge to be “reasonable” human activity in the future.

Modern businesses and technological developments thus need to be followed by appropriate regulation that will control the associated hazards and thus enable the industry to flourish. At the same time, regulation has to leave enough flexibility so that law does not restrict technological development. This development is also extremely important in the field of medicine, where technological developments bring about a special revolution in terms of medical devices, but which must also be accompanied at the right time by appropriate regulation, in order to take advantage of the benefits that new technologies bring to patients

and to avoid potential threats, either in the form of products which may harm human health or in the form of tampering with personal data and the right of individuals to confidentiality. Considering that the industry, the consumers, and patients are getting increasingly smart, smart regulatory solutions need to follow [10], establishing the right balance between safety, liability, and competition on one side and innovation and flexibility on the other. In this respect, regulatory requirements can either restrict technological development, in particular if liability for potential errors is strict or if taxation encourages human workforce, or boost it, if the standard of liability is set so that safety of computer performance is compared to the safety of certain human activity, such as driving.

In the European Union in particular, there are delicate discussions taking place on who should be competent to set the rules in this respect, Member States or EU institutions. Moreover, it is also important that this regulatory process does not bypass democratic governance principles and that industry is included in the regulatory process, as well as that self-regulation replaces legislation where possible, so that only general regulatory requirements are set by the public authorities and the market defines the technical solutions [11, 12].

In what follows, we will review social and juristic challenges of artificial intelligence in more detail, and then proceed with conclusions and guidelines as to how they might be successfully overcome.

Social Challenges

Preceding regulation and any legal action that may follow is the identification of situations where artificial intelligence is likely to be particularly challenged when it comes to making the right decision. Some situations are of course very clear-cut. A movie recommendation system should obey parental restrictions and not serve up R rated or NC-17 rated content to a child. Likewise, an autonomous vehicle should not crash into a wall for no apparent reason. But oftentimes situations are far less clear-cut, in

particular when not only the user but also others are involved.

Social dilemmas are situations where what is best for an individual is not the same, or is even at odds, with what is best for others. Already in the early 1980s, Robert Axelrod [13] set out to determine when individuals opt for the selfish option, and when they choose to cooperate and thus take into account how their actions would affect others. Of course, cooperation is a difficult proposition because it entails personal sacrifice for the benefit of others. According to Darwin's fundamental On the Origin of Species, natural selection favors the fittest and the most successful individuals, and it is therefore not at all clear why any living organism should perform an altruistic act that is costly to perform but benefits another. In Axelrod's famous tournament, the so-called tit-for-tat strategy proves to be the most successful in the iterated prisoner's dilemma game. The strategy is very simple. Cooperate first, then do whatever the opponent is doing. If the opponent was cooperative in the previous round, the strategy of tit-for-tat is to cooperate. If the opponent defected in the previous round, the strategy of tit-for-tat is to defect. This is similar to reciprocal altruism in biology.

But what about artificial intelligence, and one-off situations where the "machine" has to determine whether to act in favor of the owner (or user), or in favor of others? This was brought to an excellent point by Bonnefon et al. [14], who studied the social dilemma of autonomous vehicles. Inevitably, such vehicles will sometimes be forced to choose between two evils, such as running over pedestrians or sacrificing themselves and their passenger to save the pedestrians. The key question is how to code the algorithm to make the "right" decision in such a situation? And does the "right" decision even exist? Research found that participants in six Amazon Mechanical Turk studies approved of autonomous vehicles that sacrifice their passengers for the greater good and would like others to buy them, but they would themselves prefer to ride in autonomous vehicles that protect their passengers at all costs. Put differently, let others cooperate, i.e., sacrifice themselves for the benefit of others, but we would prefer not to.

An in essence, the same social dilemma emerges with vaccination. Old-school vaccination strategies, although admittedly easy to implement, demand that a certain fraction of the population needs to be vaccinated for herd immunity to set in. But with major progress in the structure and function of social networks [15], the same problem could be approached more systematically, by means of determining key individuals in such networks, and vaccinating them based on various metrics related to centrality, betweenness, and influence. Factors that could also be considered include temporal aspects such as traveling and commuting. A contemporary review of these developments is by Wang et al. [16]. Such advancements call for artificial intelligence, but they inevitably create a social dilemma. Will the individuals chosen for vaccination agree to this? Should they agree to being vaccinated for the good of others? Of course, the right thing to do is to agree, but the choice may nag on some of the vaccinated. Why us, why not others? The initially mentioned “old-school” strategies avoid this dilemma by essentially demanding all be vaccinated. Perhaps this is the easiest solution to the dilemma – to avoid it altogether. But easy as it may be, it neglects major progress done in many fields, including network science and digital epidemiology, and it precludes our capitalization on this progress. How many doses of vaccine could have been saved and used elsewhere for efficient immunization? How many people would not have needed to cope with some of the more adverse side effects? Would it be ethical to reward those that do get vaccinated, or even punish those that decline? We arrive at a much more complex playground of human decision-making, augmented by artificial intelligence, where a rich plethora of different strategies is at our disposal to promote cooperation [17]. But altogether, we arrive at, or rather we are faced with, the same conclusion as with the autonomous vehicles: let others cooperate and sacrifice themselves for the benefit of others, not “us.”

This is nothing if not a brutally honest outcome of a social dilemma situation involving us, humans. We are social, and we are compassionate, and we care for one another, but in rather extreme

situations, Darwin still has the best of us. It is important to understand that cooperation is the result of our evolutionary struggles for survival. As a species, we would unlikely survive if our ancestors around million years ago had not started practicing alloparental care and the provisioning for the young of others. This was likely the impetus for the evolution of remarkable other-regarding abilities of the genus Homo that we witness today [18]. Today, we are still cooperating, and on ever larger scales, to the point that we may deserve being called “SuperCooperators” [19]. Nevertheless, our societies are also still home to millions that live on the edge of existence, without shelter, without food, and without having met the most basic needs for a decent life [20].

So what can we expect from artificial intelligence in terms of managing social challenges, and in particular social dilemmas? We certainly have the ability to write algorithms that would always choose the prosocial, cooperative action. But who want to drive a car that may potentially kill you to save the lives of others? According to Bonnefon et al. [14], indeed not many of us. Hence, their conclusion, “regulating for utilitarian algorithms may paradoxically increase casualties by postponing the adoption of a safer technology.” We thus have the knowledge and the ability to program supremely altruistic machines, but we are simply too self-aware, too protective of ourselves, to then be willing to use such machines.

This in turn puts developers and engineers into a difficult position. Which is either to develop machines that are save but very few would want to buy, or to develop machines that may kill many to save one and will probably sell like honey. Nevertheless, the situation may not be as black and white, as artificial intelligence itself may learn how best to respond. Indeed, a recent review by Peysakhovich and Lerer [21] points out that, because of their ubiquity in economic and social interactions, constructing agents that can solve social dilemmas is of the outmost importance. And deep reinforcement learning is put forward as a way to enable artificial intelligence to do well in both perfect and imperfect information bilateral social dilemmas.

Well over half a century ago, Isaac Asimov, an American writer and professor of biochemistry at

Boston University, put forward the Three Laws of Robotics. First, a robot may not injure a human being or, through inaction, allow a human being to come to harm. Second, a robot must obey the orders given it by human beings except where such orders would conflict with the first law. And third, a robot must protect its own existence as long as such protection does not conflict with the first or the second law. Later on, Asimov added the fourth law, which states that a robot may not harm humanity, or, by inaction, allow humanity to come to harm. But this does not cover social dilemmas, or situations, where the machine inevitably has to select between two evils. Recently, Nagler et al. [22] proposed an extension of these laws, precisely for a world where artificial intelligence will decide about increasingly many issues, including life and death, thus inevitably facing ethical dilemmas. In a nutshell, since all humans are to be judged equally, when an ethical dilemma is met, let the chance decide. Put in an example, when an autonomous car has to decide whether to drive the passenger into a wall or overrun a pedestrian, a coin toss should be made and acted upon accordingly. Heads it's the wall, tails it's the pedestrian. No study has yet been made as to what would potential buyers of such a car make of knowing such an algorithm is embedded in the car, but it is likely safe to say that, fair as it may be, some would find it unacceptable.

Ultimately, the problems that arise when a machine's designer directs it toward a goal without thinking about whether its values are all the way aligned with humanity's, or when the machine is designed to "SuperCooperator" standards, rather harming the user than others around, we need good regulation and a prepared juristic system to tackle the challenges. This, however, leads us to a new set of challenges, namely, those that are mainly juristic.

Juristic Challenges

Considering its multifaceted character, artificial intelligence inherently touches upon a full spectrum of legal fields. Firstly, new technology raises issues concerning patentability, joint

infringement, and patent quality [23]. New relies on communication between two or more smart objects and consumers, and it is challenging whether inventors of certain types of IoT applications will be able to overcome the test for patent eligibility. Moreover, even if they obtain patents on new methods and protocols, the patents may still be very difficult to enforce against multiple infringers [23].

Furthermore, as collecting and analyzing data is progressively spreading and an increasing number of companies and health institutions have started to exploit the possibilities arising from collection and exploitation of potential data, so that added value can be created [24], this information explosion (also called "data deluge") unlocks various legal concerns that could stimulate a regulatory backlash. While it is claimed that data has become the raw material of production, and a new source of immense economic and social value [25], Big Data has been identified as "the next frontier for innovation, competition, and productivity" [26]. This is extremely relevant for the medical sector, where research is crucially dependent upon gathering sufficient amount of relevant data. On the other hand, however, open questions range from who is entitled to use this data, can data be traded, and, if so, what rules apply to this. Health data are considered particularly delicate and therefore call for special legal protection. Yet, if the rules for collecting this data are too strict, development of new medicines and health appliances might be hindered. To prevent diminishing the data economy and innovation, "smart" regulation is needed to establish a balance between beneficial uses of data and the protection of privacy, nondiscrimination, and other legally protected values. The harvesting of large data sets and the use of modern data analytics presents a clear threat for the protection of fundamental rights of European citizen, including the right to privacy [27].

Thirdly, ICT is changing the role of the consumer "from isolated to connected, from unaware to informed, from passive to active" [28]. This process is sometimes also called "digitalization" of the consumer [29], considering that people are increasingly able to use digital services. The

younger generations are grown up with digitalization and are eagerly in the forefront of adopting new technology. This could mean that the traditional presumption in consumer law that a consumer is uninformed and thus requires special legal protection no longer holds true. Nevertheless, the change is so rapid that the pre-Internet generations hardly follow the suit and new manufacturing methods bring new dangers for consumers. As the health sector greatly involves elderly generations, it is important they are included in the development and medical advances in this field while also adapting consumer law to the new challenges.

Finally, tax policy will play a very important role in the age of intelligent objects, particularly considering that human labor costs are increasing, so that it is broadly expected that automation will lead to significant job losses. As the vast majority of tax revenues are now derived from labor, firms avoid taxes by increasing automation. It is thus claimed that since robots are not good taxpayers, some forms of automation tax should be introduced to support preferences for human workers.

The focus of this section is on tort law aspects of intelligent objects, such as robots increasingly used in medicine. Tort law shifts the burden of loss from the injured party to the party who is at fault or better suited to bear the burden of the loss. Typically, a party seeking redress through tort law will ask for damages in the form of monetary compensation. Tort law aims to reduce accidents, promote fairness, provide peaceful means of dispute resolution, etc. [30].

According to the level of fault, torts fall in three general categories:

- (a) Intentional torts are wrongs that the defendant deliberately caused (e.g., intentionally hitting someone).
- (b) Negligent torts occur when the defendant's actions were unreasonably unsafe, meaning that she has failed to do what every (average) reasonable person would have done (e.g., causing an accident by speeding).
- (c) Strict (objective) liability torts do not depend on the degree of care that the defendant used; there is no review of fault on the side of the

defendant; rather, courts focus on whether harm is manifested. This form of liability is usually prescribed for making and selling defective products (products' liability).

Multifaceted character of artificial intelligence brings challenges in the field of regulating liability for damage caused by intelligent objects.

Tort Law

In relation to automated systems, various safety issues may arise, despite the fact that manufacturers and designers of robots are focused on perfecting their systems for 100% reliability and thus making liability a nonissue [31]. It can happen that robotic technology fails, either unintentionally or by design, resulting in economic loss, property damage, injury, or loss of life [32]. For some robotic systems, traditional product liability law will apply, meaning that the manufacturer will bear responsibility for a malfunctioning part; however, more difficult cases will certainly come to the courts, such as a situation, where a self-driving car appears to be doing something unsafe and the driver overrides it – was it the manufacturer's fault, or is it the individual's fault for taking over [33].

Similar difficulties may arise in relation to remotely piloted aircrafts (so-called civil drones). In the USA, a case concerning civil drones already appeared before the courts, when US Federal Aviation Administration issued an order of a civil penalty against Raphael Pirker, who in 2011, at the request of the University of Virginia, flew a drone over the campus to obtain video footage and was compensated for the flight. First instance, the court decided that a drone was not an aircraft, while the court of appeal ruled to the opposite. The cases ended in 2015 with a settlement of \$1.100.

The starting point for examining “computer-generated torts” [30] is – or at least should be – that machines are, or at least have the potential to be, substantially safer than people. Although media broadly reported on the fatality involving Tesla's autonomous driving software, it is

generally accepted that self-driving cars will cause fewer accidents than human drivers. It is stated that 94% of crashes involve human error. Moreover, medical error is one of the leading causes of death [34]. Consequently, artificial intelligence systems, like IBM's Watson, that analyze patient medical records and provide health treatment do not need to be perfect to improve safety, just better than people.

If accident reduction is in fact one of the central, if not the primary, aims of tort law, legislators should adapt standards for tort liability in case of harm caused by intelligent objects in such a way that law encourages investment in artificial intelligence and thus increases safety of humans. Most injuries people cause are evaluated under a negligence standard, where a tortfeasor is liable in case of unreasonable conduct. If her act was not below the standard of a reasonable person, the harm is thought to be pure matter of chance for which no one can be held accountable. When computers cause the same injuries, however, a strict liability standard applies, meaning that it does not matter whether someone is at fault for the harm caused or not. This distinction has financial consequences and discourages automation, because computer controlled objects incur greater liability for the producer or owner than people. Moreover, if we want to improve safety through broader use of automation, current regulation has the opposite effect.

As currently product's liability is strict, that is independent of fault, while human activity is measured according to the standard of a reasonable person, legal scholars claim that in order to incentivize automation and further improve safety, it is necessary to treat a computer tortfeasor as a person rather than a product. It is thus defended that where automation and digitalization improve safety, intelligent objects should be evaluated under a negligence standard, rather than a strict liability standard and that liability for damage would be compared to a reasonable person [30]. Additionally, when it will be proven that computers are safer than people, they could set the basis for a new standard of care for humans, so that human acts would be assessed from the perspective what a computer would have done and

how using the computer humans could avoid accidents and the consequent harm.

Nevertheless, jurists broadly defend strict liability for intelligent objects or in some respects even broader than currently foreseen, particularly in terms of the bodies involved that could be held liable – from the producer, distributor, seller, but also the telecommunication provider, when, for example, the accident was caused due to the lack of Internet connection. At the European Union level, considering that the Product Liability Directive (85/374/EEC) does not apply to intangible goods, inadequate services, careless advice, erroneous diagnostics, and flawed information are not in themselves included in this directive. It is nevertheless important that when damage is caused by a defective product, used in the provision of a service, it will be recoverable under the Product Liability Directive [35], regulating strict liability test (see also EU Court's decisions on Cases C-203/99, Veedfald, and C-495/10, Dutrueux). Many acts by robots used in medical procedures will thus come within the ambit of this Directive, including software that is stored on a tangible medium. This means that in case the consumer, whose car causes an accident due to malfunctioning software, or a patient, who suffers the wrong dosage of radiation due to a glitch in the consumer software, may bring a claim under the Product Liability Directive against the producer of software [36]. When software is supplied over the Internet (so-called non-embedded software), however, potential defects do not fall within the scope of this directive, and a specific directive on the liability of suppliers of digital content is needed.

As far as product safety regulation is concerned, Article 2(1) of Directive 2001/95 on general product safety defines the reach of the product safety regime to include any product intended for consumer use or likely to be used by consumers "including in the context of providing a service." Nevertheless, this does not cover safety of services [37]. It is hence for the EU Member States to adopt legislation setting safety standards for services, which is not the preferred solution in times of extensive technological development. Analysis of the suitability of existing safety regulations is, for example, needed in

relation to software-based product functions that can more and more be modified after delivery.

It is also essential to understand, however, that the more autonomous the systems are, the less they can be considered simple tools in the hands of other actors and that overly stringent regulation, expecting perfection instead of acceptable robot behavior, may discourage manufacturers from investing money in innovations, such as self-driving cars, drones, and automated machines [38]. Smart regulation is thus again needed, taking into account all the involved stakes.

While intelligent objects are imitating the work of humans, as well as their legal liability, the question also arises, whether robots will be entitled to sue, be sued, and also be engaged as witnesses for evidence purposes. Currently, it is not possible to sue a robot as they are considered property, just like an umbrella. Intelligent objects do not have legal identity and are not amendable to sue or be sued. If a robot causes harm, the injured party have to sue its owner or its manager. However, comparing the robots to companies, for procedural purposes companies were also not treated as separate legal entities from the human owner for a long time in history [30]. Nevertheless, over time legislators and courts abandoned the model of treating corporations solely as property and awarded them an independent artificial personality that allowed them to sue and be sued. In respect of the robots, it will thus need to be established whether they are more like an employee, a child, an animal, a subcontractor, or something else [39].

Related to this, 3D printing turns traditional service providers into manufacturers, making the relevant legislation applicable also to them. Specific regulatory challenges in this respect arise in the medical field, where 3D printing brings the ability to print replacement body parts, organs, bones, and even skin. In this situation, medical doctors and dentists provide a bundle of services – besides the ordinary patient treatment, they make a digital design of the implant and printing the implant in their offices with a 3D printer. Each device is designed and manufactured based on a patient's medical image data, which ensures a perfect fit with his unique anatomy. Low price

and high functionality 3D printed medical devices may save lives and have important consequences on the social security systems; however, the regulation needs to contemplate the risks involved and maintain patient safety standards. Under current EU Medical Devices Directive (93/42/EEC), 3D printed medical devices fall in the category of "custom-made medical devices," similar to orthopedic shoes that are not strictly regulated. In relation to 3D printed medical implants (such as prosthetic limbs, hips, or teeth), however, it is widely accepted that they require more stringent quality requirements to address the needs and potential risks [12]. Nevertheless, it seems that EU regulators are supporting the status quo, considering that the Explanatory Memorandum to the future Medical Devices Regulation states that "Manufacturers of medical devices for an individual patient, so called 'custom-made devices', must ensure that their devices are safe and perform as intended, but their regulatory burden remains low." What is thus needed to assure patients' safety is to subject the manufacturers of higher risk 3D custom printed devices to a conformity assessment and to require CE marking of the input material (in the same way as materials that are currently used for creating a dental filling). Keeping current uncertainties might lead to different national interpretations of risk related to 3D printed medical devices and a fragmentation of the internal market, thus harming both the consumers and the business.

Conclusions and Guidelines

Artificial intelligence certainly has the potential to make our lives better, especially so in medicine. It is in fact already happening, but as the adoption of any new technology, the welcoming of artificial intelligence into our lives is not without challenges and obstacles along the way. We have here reviewed some of the more obvious social and juristic challenges, for which we are nevertheless not well prepared. In particular, we have reviewed social dilemmas as traditionally demanding situations, in which we find ourselves torn between what is best for us and what is best

for others around us and for the society as a whole. It is difficult enough for us to do the right thing in such situations, and now we have to essentially build machines that will, with more or less self-training, be able to do the right thing as well. The essential question is whether we expect artificial intelligence to be prosocial, or whether we expect it to be bent on satisfying an individual, the owner, or the company of which property it is. The meme “is my driverless car allowed to kill me to save others?” brings the dilemma to the point. It is relatively easy and noble to answer yes without much thought, but who would really want a car that could potentially decide to kill you to save other strangers? Research done thus far indicates that not many, depending of course on some details as to who might the passengers be and how many others would potentially be saved. But regardless of these considerations, one of such cars is an unlikely entry on the top of any wishing list. There are of course many similar situations that have the same hallmark properties of a social dilemma, like whether or not we should be vaccinated. If a large enough fraction of a population says no, then we will lose herd immunity, and long forgotten diseases will surely return. To be vaccinated, on the other hand, is a difficult decision for some because of possible side effects of the vaccine.

Therefore, the answer to the question whether we want artificial intelligence to be prosocial or not certainly has no easy or universally valid answer. As is so often the case, it depends on the situation, and also on the juristic circumstances either decision would create.

As industry and technology are changing hastily, all the involved stakeholders have to utterly consider whether the society can adjust to this development equally fast and whether people develop the necessary technological skills. While some commentators claim that EU may adopt the legislation concerning digitizing industry too fast, since it is not yet known how exactly smart industry will develop, others call for immediate response to avoid distinct legislative activities by individual states. Robotization in many aspects makes sense and it is thus reasonable that it gets regulatory support. However, this does not mean that it is

always necessary to rush into new regulation, when amending existing legislation would suffice.

In reviewing the social and juristic challenges of artificial intelligence in medicine, we propose the following set of guidelines:

- (i) Improving the digital skills of the workforce for medical professions requires public measures with pertinent financial support.
- (ii) Strict liability for the marketing of autonomous healthcare services and medical diagnostics discourages investment in this field, thereby decreasing the potential of robotization to make these services safer and more accessible and affordable. This can be considered as the main regulatory paradox with respect to the introduction of artificial intelligence into new areas of application, including medicine.
- (iii) Patients' safety needs to be ensured by subjecting the manufacturers of higher risk 3D custom printed devices to a conformity assessment and to require CE marking of the input material (in the same way as materials that are currently used for creating a dental filling).
- (iv) Before autonomous services enter into medicine, liability issues need to be clearly set by legislation, so that it is not left to the user to search and prosecute the liable entity in courts.
- (v) Obligatory black box to record the functioning of the intelligent object and help ascertain liability in cases of potential faults.
- (vi) No fine print. The user should be informed how the artificial intelligence will react in critical situations, as well as be made accurately aware of all drawbacks, possible errors, misdiagnosis, and things that can go wrong when relying on it.

References

1. Russel S, Norvig P. Artificial intelligence: a modern approach. Pearson Education; 2013.
2. Spring M, Araujo L. Product biographies in servitization and the circular economy. Ind Mark Manag. 2017;60:126–37.

3. Kopetz H. Real-time systems: design principles for distributed embedded applications. Springer Science & Business Media; 2011.
4. Gubbi J, Buyya R, Marusic S, Palaniswami M. Internet of Things (IoT): a vision, architectural elements, and future directions. *Futur Gener Comput Syst*. 2013;29: 1645–60.
5. Chabanne H, Urien P, Susini JF, editors. RFID and the Internet of things. ISTE; 2011.
6. Wilkinson A, Dainty A, Neely A, Brax SA, Jonsson K. Developing integrated solution offerings for remote diagnostics. *Int J Oper Prod Manag*. 2009;29:539–60.
7. Stantchev V, Barnawi A, Ghulam S, Schubert J, Tamm G. Smart items, fog and cloud computing as enablers of servitization in healthcare. *Sens Transducers*. 2015;185:121–8.
8. Kryvinska N, Kaczor S, Strauss C, Greguš M. Servitization—its raise through information and communication technologies. In: International conference on exploring services science. Champions: Springer; 2014.
9. Yoo J. Embracing the machines: rationalist war and new weapons technologies. *Calif Law Rev*. 2017;105: 443–99.
10. Oettinger G. Europe's future is digital. Speech at Hannover Messe. Speech 15. 2015. p. 4772.
11. Weber RH. Internet of things – need for a new legal environment? *Comput Law Secur Rev*. 2009;25: 522–7.
12. Bräutigam P, Klindt T. Digitalisierte Wirtschaft/Industrie 4.0. Bundesverband der Deutschen Industrie; 2015.
13. Axelrod R, Hamilton WD. The evolution of cooperation. *Science*. 1981;211:1390–6.
14. Bonnefon JF, Shariff A, Rahwan I. The social dilemma of autonomous vehicles. *Science*. 2016;352:1573–6.
15. Estrada E. The structure of complex networks: theory and applications. Oxford University Press; 2012.
16. Wang Z, Bauch CT, Bhattacharyya S, d'Onofrio A, Manfredi P, Perc M, Perra N, Salathé M, Zhao D. Statistical physics of vaccination. *Phys Rep*. 2016;664:1–13.
17. Perc M, Jordan JJ, Rand DG, Wang Z, Boccaletti S, Szolnoki A. Statistical physics of human cooperation. *Phys Rep*. 2017;687:1–51.
18. Hrdy SB. Mothers and others. Harvard University Press; 2011.
19. Nowak M, Highfield R. Supercooperators: altruism, evolution, and why we need each other to succeed. Simon and Schuster; 2011.
20. Arthus-Bertrand Y. Human (movie). Bettencourt Schueller Foundation; 2015.
21. Peysakhovich A, Lerer A. Towards AI that can solve social dilemmas. In: AAAI Spring symposia 2018. Stanford University.
22. Nagler J, van den Hoven J, Helbing D. An extension of Asimov's robotics laws. In: Towards digital enlightenment. Champions: Springer; 2019.
23. Robinson WK. Economic theory, divided infringement, and enforcing interactive patents. *Fla Law Rev*. 2015;67:1961.
24. Bessis N, Dobre C, editors. Big data and internet of things: a roadmap for smart environments. Basel: Springer International Publishing; 2014.
25. Tene O, Polonetsky J. Privacy in the age of big data: a time for big decisions. *Stanford Law Rev Online*. 2011;64:63.
26. Manyika J. Big data: the next frontier for innovation, competition, and productivity. 2011. http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation
27. Lynskey O. Deconstructing data protection: the ‘added-value’ of a right to data protection in the EU legal order. *Int Compar Law Q*. 2014;63(3):569–97.
28. Prahalad CK, Ramaswamy V. Co-creating unique value with customers. *Strateg Leadersh*. 2004;32:4.
29. Mäenpää R, Korhonen JJ. Digitalization in retail: the impact on competition. Leadership in transition: the impact of digitalization on Finnish organizations. Aalto University publication series SCIENCE+ TECHNOLOGY, vol. 7. Greater Helsinki: Aalto University; 2015.
30. Abbott R. The reasonable computer: disrupting the paradigm of tort liability. *George Wash Law Rev*. 2018;86:1.
31. Kirkpatrick K. Legal issues with robots. *Commun ACM*. 2013;56(11):17–9.
32. Hilgendorf E. Robotik im Kontext von Recht und Moral. Nomos Verlagsgesellschaft; 2014.
33. Schellekens M. Self-driving cars and the chilling effect of liability law. *Comput Law Secur Rev*. 2015;31(4): 506–17.
34. Donaldson MS, Corrigan JM, Kohn LT, editors. To err is human: building a safer health system. National Academies Press; 2000.
35. Howells G, Cartwright P, Dutson S, Fawcett J, Mildred M, Willett C. The law of product liability (Butterworths Common Law). Butterworths Law; 2007.
36. Wuyts D. The product liability directive—more than two decades of defective products in Europe. *J Eur Tort Law*. 2014;5(1):1–34.
37. Weatherill S. EU consumer law and policy. Edward Elgar; 2013.
38. Richards NM, Smart WD. How should the law think about robots? In: Robot law. Edward Elgar; 2016.
39. Michalski R. How to sue a robot. *Utah Law Rev*. 2018;3.



Ethical Challenges of Integrating AI into Healthcare

9

Lisa Soleymani Lehmann

Contents

Respect for Autonomy	139
Beneficence	140
Nonmaleficence	141
Privacy	141
Safety	141
Justice	142
Impact on the Physician-Patient Relationship	142
References	143

Abstract

Artificial intelligence (AI) is revolutionizing healthcare, and with this transformative innovation comes the challenge of responsibly integrating AI into clinical care. AI has the potential to improve patient outcomes, increase the efficiency of healthcare diagnosis and treatment, and lower the cost of care. Leveraging these benefits, however, requires attention to the ethical risks raised by this new technology. In this chapter, I illuminate the primary ethical challenges of AI in healthcare and argue that in order to fully realize the potential of AI to improve individual and population health, we

need to align AI with the ethical principles of medicine. The ethical challenges posed by AI can be categorized into the four principles commonly used in healthcare ethics: respect for autonomy, beneficence, nonmaleficence, and justice [1]. Careful consideration of the implications of these principles will allow us to maximize the benefits of AI in healthcare.

Respect for Autonomy

The principle of respect for autonomy entails enabling patients to make their own decisions about which healthcare interventions they will or will not receive. Healthcare practitioners manifest respect for patients' autonomy by helping patients make informed decisions through a process of informed consent. By sharing information about the purpose, risks, benefits, and alternatives to an

L. S. Lehmann (✉)

Google, Harvard Medical School, Brigham and Women's Hospital, Boston, MA, USA

intervention and engaging patients in a process of shared decision-making, healthcare practitioners guide patients to decisions that are consistent with their goals and values. As healthcare systems and practitioners integrate AI into clinical decision-making, questions arise about whether patients should be informed that AI is being used to guide clinical decision-making and treatment. Do patients need to consent to the use of AI in clinical care?

Informed consent within the context of AI poses practical challenges of explainability to patients as the output derived from AI systems can be challenging to interpret [2]. For example, when deploying an algorithm trained to identify cancer cells, clinicians may be unable to identify the features the algorithm used to make a determination that certain cells are malignant. While there may be limited information on how an AI output is derived for a specific system, in most cases there are important features of AI that can and should be shared with patients. The details of how sensitive and specific the algorithm is for certain patients, the error rate of an algorithm, how an algorithm's accuracy compares with physicians' decisions, the safeguards put into place to detect and prevent errors, and the consequences for patients' health if the AI algorithm is biased or wrong may be more important to patients than how an algorithm arrived at a decision. A retrospective research study of AI algorithm to predict eye disease needing urgent referral performed as good as the two best-performing retinal specialists, significantly outperformed six other experts and made no clinically serious wrong decisions [3]. This type of information on AI algorithm performance may be meaningful for clinical decision-making and should be shared with patients.

The limits of an algorithm should also be disclosed to patients. As AI algorithms learn from training data, it is important for clinicians to guard against automation bias and an over-reliance on the recommendations of an AI algorithm. Like extrapolating the benefits of a medication or intervention from a clinical trial to a particular patient, it is important to understand the data on which an algorithm was trained so that

clinicians can predict and inform patients if the output will be valuable to a particular patient [4]. In addition to evidence of efficacy, it will be important to share evidence of the effectiveness of an AI algorithm in clinical practice. The full scope of benefits and risks of AI in clinical care may only be evident when algorithms are integrated into our complex clinical environments. As part of the informed consent process, clinicians should provide patients with information on the characteristics of patients whose data was used to develop the algorithm and convey the potential for an algorithm to underperform in populations on whom it has not been tested. As AI algorithms are scaled to diverse populations and we have greater confidence in their ability to improve patient outcomes, they will likely become part of the fabric of healthcare and become so commonplace that they don't require explicit informed consent.

Beneficence

Beneficence, the obligation to promote well-being, is a foundational principle of healthcare. Healthcare practitioners have a moral duty to act with the goal of benefiting our patients. Many interventions carry some potential for harm, thereby necessitating an assessment of the risk of harm and a weighing of the benefits and risks for individual patients, populations, and society. Beneficence demands that we maximize the benefits and minimize the potential harms.

AI holds the promise of improved patient health and quality of care. AI is already being used to assess patients' symptoms and provide guidance to patients on whether they should seek care in an emergency department [5]. AI applied to imaging and diagnostics is integrated into clinical care with systems that assess the degree of diabetic retinopathy in patients [6], identify malignant lesions on mammograms [7], and help clinicians diagnose dermatological conditions [8]. The application of machine learning to complex diagnostic challenges will allow for the recognition of patterns that even the best clinicians may not have recognized. AI may also allow clinicians to benefit patients by fine-tuning

screening and prevention protocols to significantly reduce the burden of disease through earlier cancer detection.

As healthcare systems integrate AI into clinical decision support systems, the potential benefits to populations including of improved quality of care, accuracy of diagnosis, efficiency, and cost of care may conflict with the potential for harm to individual patients. AI algorithms may recommend interventions that maximize aggregated benefits for a population while generating harm for a segment of individual patients with cases that the algorithm does not account for. While AI has the capacity to improve population health, it is critical to also consider how to mitigate possible risk to individuals. The risk of using AI on all populations should not be greater than what would be imposed when AI is not used.

Nonmaleficence

The benefits of AI should be considered within the context of the potential for harm. In this section, I focus on the potential for harm associated with AI secondary to privacy and safety.

Privacy

Many AI algorithms depend on large volumes of individual patient data. In the United States, the Health Insurance Portability and Accountability Act Privacy Rule (HIPAA) guides the use and disclosure of patient data [9]. In Europe, the General Data Protection Regulation (GDPR) protects EU residents' personally identifiable information. [10]. HIPPA is narrow in its scope, covering specific identifiable health information generated and held by "covered entities" and their business associates. HIPAA does not apply to de-identified health data, however, entities disclosing data should be careful to ensure that sensitive data cannot be re-pieced together.

GDPR is an entirely different regulatory regime with distinct requirements that must be met to process sensitive personal data, including data concerning health. For example, data

concerning health may only be processed under certain circumstances, such as the individual providing explicit consent or processing that is necessary for direct patient care, public health, scientific, historical or statistical research purposes [11]. In most instances, individuals must consent to their personal data being collected and used for automated decision-making that significantly or legally affects them, and individuals have the right to obtain human intervention and contest decisions based on AI [12].

Privacy is linked to digital agency, that is, control over access to and use of an individual's personal information [13]. Patients are increasingly demanding greater transparency over what personal health data is collected for, who has access to their personal health data and how it is used.

Natural language processing (NLP) applications are also being integrated into clinical settings and raise privacy considerations. As healthcare systems try to enhance efficiency with the use of chatbots for patient communication and improving mental health, patients are expressing concerns about electronic sharing of confidential information [14].

Safety

Researchers have increasingly recognized AI algorithms need broad validation to be able to make predictions in varying populations. The data used to train an algorithm has significant implications on the output of the model. Subtle differences between data used to train an algorithm and real world data encountered during clinical deployment can result in algorithms making erroneous predictions that may be difficult to identify. For example, a sepsis prediction model deployed in hundreds of hospitals was found to perform worse than indicated by initial studies. This example underscores the need for external validation of AI algorithms before widespread use [15, 16].

A recent study of AI predictions of hospital-acquired infections found that the variables

associated with risk at one hospital were protective at another hospital [17]. As a result of using inaccurate synthetic data, as opposed to real patient data an AI algorithm generated erroneous recommendations for cancer treatment [18]. These examples underscore the importance of broad testing of AI algorithms in diverse patient populations and adopting safeguards, such as using AI as an adjunct to clinical decision-making, until we have confidence in a model's predictions. An AI algorithm that is not generalizable across populations can pose a threat to clinical decision-making and patient safety and, similar to other clinical interventions should be tested in large groups of diverse patient populations [19, 20]. Our ethical duty to do no harm requires the development and integration of proactive and responsive quality management systems alongside the deployment of AI in healthcare.

Justice

One of the most significant ethical concerns with the integration of AI into healthcare is the potential to introduce and amplify bias based on race, sex, insurance status, and healthcare usage [21, 22]. Such bias may only emerge when an algorithm is employed in diverse populations and can exacerbate health inequalities [23]. The ethical principle of justice in healthcare impels us to prioritize fairness in the provision of healthcare. It would be unjust to give one group access to a diagnostic or therapeutic intervention and withhold it from another group who suffers with the same illness.

Despite the importance of justice in healthcare, bias has crept into clinical prediction models. In a recent study of the use of an AI algorithm used to predict risk and guide decisions for referral to high-risk care management programs used for roughly 200 million patients, investigators found that Black patients assigned the same level of risk by the algorithm were sicker than white patients [23]. Racial bias in the algorithm reduced the number of Black patients identified for extra care

by more than half. This error was thought to occur because the algorithm used healthcare costs as a proxy for health needs. Because less money was spent on Black patients with the same level of need, the algorithm falsely concluded that Black patients were healthier than equally sick white patients. The prediction model used the cost of care as a proxy for healthcare needs, and those who created the model did not realize this seemingly minor decision could have such far-reaching consequences for patient care. This research exemplifies the need to rigorously evaluate models prior to integrating them into clinical care and the importance of a socially conscious approach to the development of AI algorithms so that we don't encode and amplify inequities in healthcare.

Another factor driving bias in AI is the underrepresentation of diverse groups in training and testing datasets used to develop and validate AI models [25, 26]. The benefits of AI in healthcare are going to be limited by the scarcity of data that adequately represent individuals of varying age, sex, race, ethnicity, and environments [27]. Healthcare practitioners can help address this challenge by increasing awareness among underrepresented populations of the value of data sharing and by building trust through transparent and inclusive communication.

Impact on the Physician-Patient Relationship

AI is likely to change the nature of the physician-patient relationship. Patients look to their physicians as their fiduciary who will advocate for their best interests. With greater reliance on the output of AI algorithms, healthcare practitioners may feel less empowered to challenge an AI decision or advocate for a particular patient. The sharing of patients' data without their meaningful and explicit consent may also undermine patients' trust in their healthcare practitioners and the profession of medicine. Patients may express understandable anxiety about the role of difficult to explain algorithms influencing healthcare

decisions. Clinicians may mitigate these concerns by explaining how the AI was integrated into their decision making and recommendation. There is a risk that an overreliance on the technology will diminish physicians' clinical skills and that our constant push for greater efficiency will allow the outputs of AI systems to be directly shared with patients without appropriate context and interpretation provided by clinicians. Some patients may have a preference for receiving clinical information and treatment recommendations from an empathic physician. Clinicians have an important role to play in educating patients about the risks and benefits of AI in healthcare.

AI has the potential to radically change the practice of medicine through the application to diagnosis and treatment. Aligning the integration of AI with the ethical principles of our profession, including respect for patient autonomy, beneficence, nonmaleficence, and justice, will accelerate the responsible adoption of AI into daily clinical practice.

References

1. Beauchamp TL, Childress JF. Principles of biomedical ethics. 5th ed. New York: Oxford University Press; 2001.
2. Castelvecchi D. Can we open the black box of AI? *Nature*. 2016;538(7623):20–3.
3. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. 2018;24:134250. <https://doi.org/10.1038/s41591-018-0107-6>.
4. Pellegrino ED, Thomasma DC. The conflict between autonomy and beneficence in medical ethics: proposal for a resolution. *J Contemp Health Law Policy*. 1987;3:23.
5. Ada. Your personal health guide. <https://ada.com>; 2021.
6. van der Heijden AA, Abramoff MD, Verbraak F, van Hecke MV, Liem A, Nijpels G. Validation of automated screening for referable diabetic retinopathy with the IDx-DR device in the Hoorn Diabetes Care System. *Acta Ophthalmol*. 2018;96(1):63–8. <https://doi.org/10.1111/aos.13613>.
7. Salim M, Wählén E, Dembrower K, et al. External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms. *JAMA Oncol*. 2020;6(10):1581–1588. <https://doi.org/10.1001/jamaoncol.2020.3321>
8. Brinker TJ, Hekler A, Enk AH, Berking C, Haferkamp S, Hauschild A, et al. Deep neural networks are superior to dermatologists in melanoma image classification. *Eur J Cancer*. 2019;119:11–7.
9. <https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/combined-regulation-text/index.html>
10. <https://eur-lex.europa.eu/eli/reg/2016/679/oj>
11. GDPR Article 9 (2) a.
12. GDPR Article 22.
13. Aitken M, de St Jorre J, Pagliari C, et al. Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies. *BMC Med Ethics* 2016;17(1):73.
14. Utermohlen K. Four robotic process automation (RPA) applications in the healthcare industry. *Medium*, 2018. <https://medium.com/@karl.utmohlen/4-robotic-process-automation-rpa-applications-in-the-healthcare-industry-4d449b24b613>
15. Amodei D, Olah C, Steinhardt J. Concrete problems in AI safety. *arXiv [cs.AI]*. 06565. 2016.
16. Wong A, Otles E, Donnelly JP, et al. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Intern Med*. 2021;181(8):1065–1070.
17. Oh J, et al. A generalizable, data-driven approach to predict daily risk of *Clostridium difficile* infection at two large academic health centers. *Infect Control Hosp Epidemiol*. 2018;39:425–33.
18. Hernandez D, Greenwald T. IBM has a Watson dilemma. *The Wall Street Journal*. August 11, 2018. www.wsj.com/articles/ibm-bet-billions-that-watson-could-improve-cancer-treatment-it-hasnt-worked-1533961147
19. Char DS, Shah NH, Magnus D. Implementing machine learning in health care – addressing ethical challenges. *N Engl J Med*. 2018;378(11):981–3.
20. Cabitzia F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *JAMA*. 2017;318(6):517–8.
21. Ferryman K, Winn RA. Artificial intelligence can entrench disparities: here's what we must do. *The Cancer Letter*, Nov 2016.
22. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, Jung K, Heller K, Kale D, Saeed M, et al. Do no harm: a roadmap for responsible machine learning for healthcare. *Nat Med*. 2019;25(9):1337–40.
23. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med*. 2018;169(12):866–72. <https://doi.org/10.7326/M18-1990>.
24. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366 (6464):447–53. <https://doi.org/10.1126/science.aax2342>.
25. Ada Lovelace Institute. Black data matters: how missing data undermines equitable societies. <https://www.adalovelace.com/black-data-matters/>

- adalovelaceinstitute.org/black-data-matters-how-missing-data-undermines-equitable-societies
26. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. *JAMA Dermatol.* 2018;154(11):1247–8. <https://doi.org/10.1001/jamadermatol.2018.2348>.
27. Ibrahim H, Liu X, Zariffa N, Morris AD, Denniston AK. Health data poverty: an assailable barrier to equitable digital health care. *Lancet Digit Health.* 2021;3(4):e260–5.



Artificial Intelligence in Medicine and Privacy Preservation

10

Alexander Ziller, Jonathan Passerat-Palmbach, Andrew Trask,
Rickmer Braren, Daniel Rueckert, and Georgios Kaassis

Contents

Introduction	146
General Considerations and Current Technical Standards	147
Anonymization, Pseudonymization, and k -Anonymity	148
Considerations for Specific Dataset Types	148
The Requirement for Next-Generation Privacy-Preserving Techniques	149
Federated Learning	149
Technical Framework	149
Challenges in Federated Learning	149
Attacks Against Federated Learning Systems	150
Applications of Federated Learning	150
Differential Privacy	151
Properties	151
Implementation	151
Sensitivity and Privacy Budget	151
Challenges	152
Applications	152

A. Ziller · G. Kaassis (✉)

Institute for Diagnostic and Interventional Radiology,
School of Medicine, Technical, University of Munich,
Munich, Germany

Institute for Artificial Intelligence in Medicine and
Healthcare, School of Medicine and Department of
Informatics, Technical University of Munich, Munich,
Germany

OpenMined, Oxford, UK
e-mail: alex.ziller@tum.de; g.kaassis@tum.de

J. Passerat-Palmbach
OpenMined, Oxford, UK

Department of Computing, Imperial College London,
London, UK

Consensys Health, New York, NY, USA
e-mail: j.passerat-palmbach@imperial.ac.uk

A. Trask

OpenMined, Oxford, UK

Department of Computer Science, University of Oxford,
Oxford, UK

e-mail: andrew@openmined.org

R. Braren

Institute for Diagnostic and Interventional Radiology,
School of Medicine, Technical, University of Munich,
Munich, Germany

e-mail: rbraren@tum.de

D. Rueckert

Institute for Artificial Intelligence in Medicine and
Healthcare, School of Medicine and Department of
Informatics, Technical University of Munich, Munich,
Germany

Department of Computing, Imperial College London,
London, UK
e-mail: d.rueckert@tum.de

Homomorphic Encryption and Secure Multi-Party Computation	152
Homomorphic Encryption	152
Secure Multi-Party Computation	153
Trusted Execution Environments	154
Outlook	154
References	155

Abstract

The widespread applicability of medical artificial intelligence systems hinges on their development and validation on large, diverse, and representative datasets. So far, such datasets have only been able to be assembled through multi-institutional data sharing and aggregation. Such practices are however associated with legal, ethical, and technical challenges and scale poorly to multinational efforts. They furthermore potentially infringe on data ownership and complicate the enforcement of data governance measures. Privacy-preserving machine learning offers solutions to these challenges by implementing techniques for the decentralized training of algorithms on datasets without requiring direct access to the data or by offering guarantees of privacy protection during training and algorithm inference. This chapter presents the core techniques of secure and private artificial intelligence, which can serve to enable the training of algorithms on larger datasets and their provision to more people under provable assurances of privacy and ownership protection.

Introduction

Medical artificial intelligence (AI) research has, over the past few years, led to several successful clinical applications such as individualized cancer diagnostics, tumor subtype characterization, patient outcome prediction, and survival or therapy response risk assessment [1–9]. It is to be expected that AI systems will become part of clinical routine in the near future and offer decision support to physicians in a variety of

domains. The common denominator of such applications is the requirement to train AI algorithms on large and diverse datasets. Data is required to be representative of real-world patient attributes including edge cases, and – optimally – be highly diverse with regard to patient factors (such as gender, ethnicity, or age) to be fair and unbiased, but also with regard to different imaging system vendors to be applicable in real-world scenarios. Such datasets, ideally representing all relevant clinical pathologies encountered in patient care, can usually not be procured by single sites alone; rather, international confederations of healthcare providers are required. The realization of such collaborations is so far based primarily on centralized data aggregation and sharing. However, this practice is problematic in several regards: Legislation regarding data protection and sharing is highly heterogeneous across different countries. The US Health Insurance Portability and Accountability Act (HIPAA) [10] imposes strict rules on the use of personal health data and requires authentication, authorization and accountability. The European Union's General Data Protection Regulation (GDPR) [11] stipulates rights for data access, rectification, and deletion (*forgetting*) of stored data, the restriction of processing, and a strictly regulated view on data portability and sharing. In the context of AI-based processing, this regulatory landscape raises complex considerations, such as verifiable algorithmic interpretability to realize rights to explanation, as well as questions such as ownership of algorithms and data, and the enforcement of granular governance schemes. Such considerations are beyond the scope of this chapter, but will without doubt require multidisciplinary scientific, political,

and societal debate. Crucially, such debate must address fundamental ethical principles, such as transparency, justice and fairness, non-maleficence, and responsibility [12, 13].

Beyond such considerations, applications including medical data are critical due to the high value of the data [14]. The release of health information can lead to severe discrimination by employers or health insurance providers, with allegations pointing to systematic profiling of high health-risk individuals and business models involving the sale of de-anonymized medical records to insurance companies to mitigate financial risks [15]. Such practices also highlight the function of personal information within an evolving data economy, in which data assumes the role of a valuable production asset. An increasingly privacy-aware public, conscious of this value, is thus demanding *single-use-accountability*, that is, the ability to at any time rescind the permission about the storage, transmission, and/or utilization of their personal data and the guarantee that the data is used only for an explicitly designated purpose. Similar considerations are raised by algorithm owners, who wish to protect their AI models, representing both production value and intellectual property assets, from theft, misuse, or unauthorized exploitation. It thus becomes clear that medical AI applications are developing within a complex, multifaceted field involving algorithm developers, imaging equipment vendors, physicians, and patients, each with their own motives, objectives, and – potentially conflicting – interests. This complexity, lack of transparency regarding motives and the nature of digital data being more accessible, easier to exchange and distribute [16] than ever, are threatening the inviolability of data privacy. While the importance of privacy is undoubted, its precise definition is also evolving. Historical definitions such as the *right to be let alone* from American law scholars in 1890 [17] are no longer applicable in the context of AI, as they could not foresee the technical advances including digitization, the Internet, or data science. Kaassis et al. define privacy as *the ability to retain full control and secrecy over one's personal information* [18]. The *control* aspect is further emphasized by

definitions such as Nissenbaum et al. [19], who eschew notions of privacy based primarily on obfuscating data, instead defining it as *contextual integrity*, that is, the purpose-specific flow of data within a larger context of societal and ethical norms. The concept of privacy should also be delimited from *confidentiality*, the legal or ethical duty to maintain secrecy over specific information.

AI in medicine represents a quintessential conflict between privacy and utility (*privacy-utility-tradeoff*), that is, between the **protection** of personal data and AI algorithms and the **utilization** of personal data for AI algorithm training. The reconciliation of these requirements is the domain of *secure and private artificial intelligence*. It provides a set of techniques for ascertaining the protection of data (*training data, model input, and model output privacy*) and algorithms (*model privacy*). The provision of these guarantees alongside auditability and integrity assurances via technically verifiable methods is termed “structured transparency” [18]. This new and evolving discipline is embedded within the larger field of *trustworthy artificial intelligence*, which includes the study of fairness, robustness, causality, interpretability, security, privacy, verifiability, accountability, and auditability. The current chapter will motivate the requirement for next-generation privacy-preserving methods, present their technical realization by summarizing the concepts of *federated learning, differential privacy, homomorphic encryption, secure multi-party computation, and trusted execution environments*, explain their current technical limitations, and sketch a scenario of how their interplay enables large-scale decentralized and privacy-preserving machine learning in medicine.

General Considerations and Current Technical Standards

Prior to a description of specific techniques, the general idea of a *threat model* is introduced. All techniques described in this chapter unfold within this conceptual framework defining the level of trust between participants. Scenarios can

range from *fully trusted* settings, in which all participants adhere to protocols without attempting to extract any information to *honest-but-curious*, in which participants do not actively undermine the protocol, however try to extract as much information as possible to *dishonest*, in which participants actively attempt to undermine or subvert protocols or the entire system to gain information or cause harm. Such dishonest participants are also referred to as *adversaries*. A threat model specifies the theoretical considerations required and the level of protection offered by a system under a given level of trust.

Current technical standards of privacy preservation revolve mainly around de-identification, that is, the removal or obfuscation of personally identifiable data from datasets. As mentioned in the introduction, the requirement for large and diverse datasets is currently addressed by centralized data sharing of such de-identified data in a majority of cases. Three main techniques for de-identification exist, which are described in the following section.

Anonymization, Pseudonymization, and k -Anonymity

Anonymization refers to the removal of all identifiable patient data entries, for example, patient name or birth date. Pseudonymization is similar but does not rely on removing these entries, but rather on replacing or altering them using randomly generated identifiers in a non-recoverable fashion. Usually, the pseudonymization provider retains a linkage record, or so-called *look-up table* to later be able to re-identify certain individuals if a specific question arises, or the dataset is modified. An evolution of these techniques *k -anonymity* [20] provides a theoretical guarantee that an anonymized data sample cannot be distinguished from $k - 1$ other samples in the corresponding dataset. It is of note that stronger notions of anonymity correspond to a decrease in information content of the dataset, shifting the privacy-utility tradeoff toward privacy. Although representing the current standard, all aforementioned methods are prone to *catastrophically degrade* in the

presence of side information. Catastrophic degradation refers to the full reidentification of individuals contained in the dataset, also referred to *blatant non-privacy*. Side information can be thought of as any (partially) non-de-identified dataset containing at least one identifier in common with the de-identified dataset. Such side information can give rise to a class of attacks termed *linkage attacks* [21]. They are based on the usage of a non-anonymized dataset containing an intersection of patients with the attacked dataset and finding similarities that are sufficiently remarkable such that, with a high probability, two entries are from the exact same patient. A prominent example of a linkage attack was the de-anonymization of the *Netflix prize* dataset by Narayanan et al. [22] in 2008. The dataset contained 500,000 anonymized movie ratings of Netflix users and was linked to profiles of known users of the IMDb dataset, who were identifiable by name. The same technique was then used to recover information about political preferences and other sensitive information about the users. Similarly, the de-anonymization of New York City taxi data [23] was demonstrated, in which the authors were able to concretely reidentify 91% of the contained data samples. Modern linkage attacks are estimated to be able to concretely identify 99.98% of the American population using 15 demographic attributes [24].

Considerations for Specific Dataset Types

The abovementioned de-identification methods are unusable for certain types of data containing a high inherent density of information and requiring more sophisticated methods for anonymization. A typical example is medical imaging data, which – even with no additional information available – contains obviously identifiable physical properties of the patient, such as the face or specific pathological alterations visible in the image. It has been demonstrated that the combination of three-dimensional rendering techniques applied to CT and/or MRI scans of the head with facial recognition software is able to

reidentify patients contained in medical imaging datasets from for example, social media images [25, 26]. One approach to prevent such attacks is the so-called *de-facing* of imaging datasets, that is, processing of the image dataset such that no identifiable structure of the face remains [27]. However, such a process may be prone to technical failure and thus requires quality control. Furthermore, certain imaging characteristics are required for diagnosis, such as the depiction of reproductive organs, which are nevertheless sufficient to identify a patient's biological gender, which might still represent data leakage. Lastly, it is unclear whether other image-derived characteristics, such as body shape or specific pathologies, might also be able to lead to concrete reidentification.

The Requirement for Next-Generation Privacy-Preserving Techniques

It becomes clear that the abovementioned current-generation de-identification techniques all suffer from notable drawbacks: They provide privacy guarantees only within a very restricted scope, are prone to degradation in the presence of side information and – for the most part – do not provide a quantifiable notion of privacy. The common denominator of next-generation privacy-preserving machine learning (PPML) techniques, which will be presented next, is their motivation by information-theoretic notions of privacy, and their provision of quantifiable guarantees of privacy or resilience to given attacks under given threat models.

Federated Learning

A conceptually simple method to eschew centralized data sharing, and thus restore both sovereignty of the data owners and the ability to enforce arbitrary governance schemes over datasets is to no longer share data, but instead send algorithms to the sites at which data is kept for training models locally. The transmission of model copies to the sites of data residence for

local training via distributed computation techniques is termed *Federated Learning* (FL) [28].

Technical Framework

FL in medicine is usually performed in the setting of *cross-silo federated learning* [29]. Here, each site (*data silo*) contains significantly more than one data sample (usually an entire dataset), on which the distributed algorithms are trained. For training, a network is established in which each site (or *node*) represents one client containing data within a *hub-and-spoke topology*, and a central instance (*hub* or *server*) plays the role of training coordinator [29]. So-called *peer-to-peer* approaches, in which the model is sent in a chain between training institutions, also exist; however, they have been shown to yield lower performance [30]. In the hub-and-spoke topology, the AI algorithm is distributed to the nodes, where a number of training iterations is performed. Then, the model parameters are *aggregated*, usually by taking their *federated average* and the updated models redistributed to the nodes. At the end of training, all nodes and the hub contain the trained algorithm without any data ever having left the nodes. Several studies [30, 31] have shown that the performance of centralized and federated learning performed in this fashion can be on par. However, federated model training introduces a series of challenges over its centralized counterpart.

Challenges in Federated Learning

The centralized aggregation of datasets provides several benefits with respect to dataset curation. Thus, quality control of datasets is facilitated, data harmonization steps can be conducted on the entire dataset and corrections to the dataset can be easily undertaken. Furthermore, access to aggregated data can help data curators ascertain fairness and representativeness of the training data and remove unwanted samples which would lead to deterioration of the algorithm's performance. These factors are substantially complicated in the

federated setting, in which the coordination of multiple data curators is required.

Furthermore, the specific statistical data distribution on the nodes can impact model training. For example, imbalanced data with respect to certain attributes (e.g., overrepresentation of certain classes), extreme statistical characteristics of the data (*outliers*), incorrect data labels (*label noise*), or low quality/unrepresentative data can all lead to poor performance of the federated model. It is a common modeling assumption that data on a node represents a representative sample drawn at random from a population (in the case of FL, the union of all datasets). Under certain circumstances, the data on the nodes might violate this assumption to such a degree that training a model locally (*personalization*) instead of contributing to or utilizing the collaboratively trained model might be preferable.

Moreover, methods to evaluate data quality in the federated setting without direct access to the data are still under investigation. In addition, FL raises questions of reimbursement and model ownership. Reimbursement of data providers can, for example, be calculated on the grounds of how much the final model profits from this data, which is dependent on a number of factors such as dataset size, quality, and diversity. Recently, game-theoretic approaches have been introduced to assist with such quantification [32]. Ownership of the final model in the setting of multiple stakeholders including the algorithm developers, data contributors, imaging equipment vendors, health insurance companies, and patients is also a complicated question requiring further interdisciplinary discussion.

From a technical point of view, federated model training introduces considerable network overhead as model parameters have to be continuously sent over the network, which, considering the size of modern deep neural network architectures with several million parameters, is a considerable amount of traffic and can represent a significant bottleneck for training time. Efficiency-centered methods, such as model compression, quantization, or the design of specific architectures adapted to distributed computation workflows, will therefore be required to address

these challenges. For a comprehensive literature review on FL challenges, see [29].

Attacks Against Federated Learning Systems

Last but not least, federated learning systems are in themselves not fully privacy-preserving. They are vulnerable to attacks that extract information about data from the model weights or gradient updates [33, 34]. Such attacks include, for example, *model inversion* or *dataset/feature reconstruction attacks*, which aim at reconstructing patient attributes or entire datasets from the model weights and derived attacks, such as *membership inference*, aimed at concluding whether an individual was present in a certain dataset and thus likely had a specific attribute. Such *privacy-focused* attacks mean that federated learning must be combined with further techniques, discussed below, to be fully privacy-preserving. Furthermore, attacks also exist against the federated learning system itself, termed *utility-focused* attacks. For example, under certain threat models, malicious participants can perform *model poisoning attacks* [35] by which low-quality data or data specifically designed to manipulate the algorithm (adversarial data) is deliberately introduced. Techniques such as *federated adversarial learning* aim to harden federated learning systems against such adversarial changes to the input space [36]. Finally, unencrypted access to models sent over the network can also give rise to model theft, thus infringing on the intellectual property of the model owners.

Applications of Federated Learning

Federated learning has already yielded several promising applications. The European Union has funded large-scale academic and mixed consortia including academia and industry, to advance federated machine learning, such as the MELLODDY consortium [37], which announced the establishment of a large-scale FL platform for drug discovery in September 2020. The

FeatureCloud consortium, also supported by the European Union, is developing FL tools which have been utilized in genome-wide association studies [38]. In the medical imaging field, mixed consortia such as the London Medical Imaging & AI Centre for Value Based Healthcare, including companies such as NVIDIA have recently showcased federated learning for brain tumor segmentation [39, 40]. Finally, during the COVID-19 pandemic, federated learning was proposed for collaborative training of predictive models on electronic health records to improve outcome risk prediction [41].

Differential Privacy

As mentioned previously, a key drawback of de-identification methods such as anonymization is—generally their lack of quantifiability regarding the specific *amount* of privacy protection they provide. Differential privacy (DP) is an attribute of an algorithm and provides a method to quantify the change in the output of the algorithm caused by the addition or deletion of a row in a dataset using information-theoretic measures [42]. Furthermore, it permits the targeted alteration of either the data or the algorithm training procedure to assure that the individual’s privacy loss in the worst case remains bounded by some quantifiable limit. Dwork et al. formulate the guarantees offered by DP as a promise: “You will not be affected, adversely or otherwise, by allowing your data to be used in any study or analysis, no matter what other studies, data sets, or information sources, are available” [43]; equivalently, the individual retains *plausible deniability* of the fact that their data was used for a certain analysis.

Properties

DP has several useful properties, such as the fact that DP algorithms, when used in conjunction, also yield a DP algorithm (*closure under composition*). Furthermore, DP offers resilience to *post-processing*, that is, privacy is retained no matter what further algorithmic processing is applied to

the output of a DP algorithm. Moreover, its guarantees generalize to datasets differing by more than one individual (*group privacy*) [43, 44]. The predictable behavior of DP in these scenarios is sometimes termed *graceful degradation* to delimit it from the catastrophic degradation of other techniques discussed above. As originally defined, DP is now usually termed *epsilon*-DP. However, due to the overly pessimistic outlook on privacy loss under this definition, newer definitions exist, such as *epsilon/delta*-DP or *R&Qc niy*-DP, which are, however, beyond the scope of this chapter [43, 44].

Implementation

DP is typically realized by targeted perturbation of some part of the data/algorith pipeline, typically by random noise injection. For instance, noise can be added to the data before processing (*input-DP*) or to the result of the analysis (*output-DP*) [45]. However, this technique is only effective in a limited set of scenarios. For instance, when considering imaging data, additional noise to the input does not change the inherent physical properties of the image and thus does not contribute to DP; hence other techniques, such as stochastic output perturbations [46], the generation of DP synthetic data [47], or differentially private algorithm ensembles [48], are utilized. Noise perturbation can also be applied to the training of algorithms using techniques such as *differentially private stochastic gradient descent*, customized for training neural networks in a differentially private manner [49].

Sensitivity and Privacy Budget

A central concept in DP is the notion of *sensitivity*, capturing the impact of an interaction with the dataset (*query*) on privacy guarantees [42], from which the amount of perturbation required for a certain level of privacy can be assessed. As an extension of this consideration, the concept of a *privacy budget* is derived. Such a budget can be assigned by the data owner or curator to third

parties wishing to interact with or process the dataset in order to monitor and limit their interaction to a certain scope. After the *budget* is exhausted, no further interaction is allowed and thus, repeated queries (e.g., *averaging attacks* [50]) are prevented from leading to a neutralization of the DP mechanism.

Challenges

It should be noted that DP is conceptually similar to other techniques leading to reduced reliance of algorithm training on individual data samples with the aim of improved generalization and robustness and reduced overfitting, such as regularization [51]. This highlights that privacy and robustness are not incompatible (but rather mutually reinforcing) concepts. Nevertheless, strong DP guarantees can dramatically shift the privacy-utility tradeoff away from high algorithm performance. Furthermore, new considerations are introduced, such as the definition of how *acceptable* privacy should be quantitatively defined. This difficulty of definition leads to challenges, for example, when attempting to assign privacy budgets. Moreover, although in the general case, DP protects the individual from privacy loss due to their inclusion in a certain analysis, the results of the analysis can still lead to negative consequences for the individual. For example, although it might remain hidden that an individual who is a smoker partook in a study to answer the question whether smoking causes cancer, the result of the study (i.e., that smoking indeed causes cancer) might nevertheless lead to an increased insurance premium for the individual.

Applications

DP is a flexible and powerful concept and can be applied at all stages of interaction between algorithms and data. For instance, local DP can be utilized to protect the privacy of sensitive data collected from wearable devices measuring health-related parameters [52]. DP at analysis or publication time has for instance been showcased

in genomics studies [53, 54], electronic health record studies [55], or for enabling the privacy-preserving dissemination of analysis results [56]. More recently, DP has been combined with federated learning for large-scale biomedical dataset analysis [57]. Further surveys and reviews on the utilization of DP in health care can be found in [58–60].

Homomorphic Encryption and Secure Multi-Party Computation

An idealized method to provide end-to-end security and privacy to machine learning systems is to encrypt both the data and the algorithms. Such a system would offer the perfect secrecy guarantees of cryptography in the setting of, for example, *diagnosis as a service*. For instance, a patient can encrypt their data and send it in encrypted form to a service provider where an encrypted algorithm resides. The algorithm can then perform inference on the data and send back an encrypted diagnosis, which can only be decrypted by the patient. The paradigm described is termed *end-to-end encryption* and can enable data to be processed by an untrusted third party, such as a model owner, without revealing sensitive information to them and likewise allow the algorithm owner to encrypt their model and prevent its theft or misuse. More generally, such schemes can be utilized to enforce *single-use accountability*, that is, the guarantee that a system or good (such as data) will only be utilized for a designated purpose.

Two main cryptographic approaches will be covered, which allow the training of AI models or their use for inference, homomorphic encryption and secure multi-party computation.

Homomorphic Encryption

Homomorphic encryption (HE) can be considered an extension to existent cryptographic protocols such as public key cryptography with the added guarantee of enabling structure-preserving mathematical computations to be performed on encrypted data. This means that encrypted data

can be processed as if it were unencrypted and the results of applying the algorithm to encrypted data and then decrypting it are the same as if unencrypted data had been processed.

Not all HE protocols however allow arbitrary operations to be performed. Older HE protocols for instance support only addition and/or multiplication operations, with newer, *fully homomorphic encryption* schemes only having been introduced more recently. Such schemes permit the training of encrypted neural networks over encrypted data [61] enabling the *end-to-end* privacy preserving attributes outlined above. Other work has focused on encrypting neural networks for performing inference in a *machine learning as a service* setting [62]. However, some fully HE schemes rely on approximate calculations. For example, the CKKS Scheme [63] introduces noise to the computational process. The application of HE to neural networks typically relies on approximations of, for example, activation functions to a certain degree of precision. Both can diminish the accuracy of the models. Furthermore, HE introduces substantial computational overhead, in some cases introducing latency which lies in orders of magnitude above unencrypted computations.

Applications

HE applications relying on efficient machine instructions [64] or theoretical advances [65] have recently been proposed to limit this impact. It is to be expected that dedicated cryptographic processing hardware will further reduce latency and computational overhead and render HE applications for fully encrypted machine learning feasible in the near future. Recent biomedical applications of HE include the utilization of *leveled HE* for genome-wide association studies [66]. Moreover, HE can be employed for carrying out individual encrypted operations such as the secure encrypted aggregation of models in federated learning [67]. For further information on HE refer to [68].

Secure Multi-Party Computation

Secure multi-party computation (SMPC) is a set of protocols for the obfuscation of information

by distributing it to multiple parties while enabling the joint application of computations without sharing information about input or output [69, 70]. From an information-theoretic point of view, the information gain of any single member in an SMPC scheme is equal to a situation in which a perfectly trustworthy third party has access to all data and performs the computations, thus guaranteeing privacy. Furthermore, SMPC protocols can be designed in a fashion that requires a quorum of participants to reconstruct the output, while the collusion between fewer than the minimum required number of participants does not offer any information whatsoever. These attributes can be used to guarantee correctness and shield against adversaries within the system. Thus, SMPC can be used under the premise (threat model) that not all participants in a network are trustworthy. Although protocols for *two-party* computation exist, the authors only describe protocols for more than two participants here. An example for such a protocol is *Shamir's secret sharing* [71]. Each data sample (*secret*) is split into several *shares* and distributed among all parties. Collectively, the output of some function (e.g., addition) can be calculated and made public upon agreement. Compared to HE, whose overhead is primarily computational, SMPC entails high communication overhead as most or all parties are typically required to be online at all times and the protocol's design relies on the exchange of large integers and additionally scales with the number of participants. Newer, communication-efficient protocols aim to minimize this penalty [72].

Applications

SMPC, like HE, can be used to encrypt both data and AI models and can thus be used in the setting of end-to-end encrypted services. Its utilization has been proposed for collaborative analytics across healthcare data silos [73] and wireless sensor networks [74]. Recent pilots also implement SMPC for optimizing workflows in hospitals via privacy-preserving location tracking of personnel and in cross-sectional population studies [75]. In the setting of the SARS2-Coronavirus pandemic,

SMPC was furthermore utilized for privacy-preserving contact tracing [76].

Trusted Execution Environments

Trusted execution environments (TEEs) are an isolated memory area created by specific machine instructions. The memory area created is called *enclave* and represents a processing environment where computations can be carried out securely [77] without any other process on the system being able to access the encrypted memory area. It provides guarantees for authenticity and confidentiality of the code upon startup, with some implementations also ensuring the integrity of runtime states. These approaches can be used in different scopes to either just protect small secrets, such as private keys, or provide system-wide encryption for secure and private computing.

In secure and private AI, TEEs can be used either in conjunction, or as a drop-in replacement to pure software cryptography methods like SMPC and HE. They have indeed been shown to deliver performance suitable for real-time deployment of ML models behind secure inference APIs [78]. At the time of writing, they represent a more developer-friendly solution to implement PPML solution than SMPC and HE thanks to abstraction frameworks such as *Graphene* [79] which allows transforming container images into *enclaves* ready for execution on appropriate hardware (in this case, Intel *software guard extensions* (SGX). TEEs are however notoriously susceptible to *side-channel attacks* [80]. Furthermore, it is obviously harder to patch vulnerabilities in hardware modules once they are in commercial use than to provide software updates, as would be possible when attacks or vulnerabilities against pure-software cryptography are detected.

The risk of vendor lock-in for TEEs can be regarded as low, seeing as standards such as the *Open Enclave* software development toolkit (<https://github.com/openenclave/openenclave>) have recently emerged to allow hardware-agnostic implementations. However, it is worth noting that Intel SGX is the most commonly solution as of today, which recalls the quasi-monopoly situation

in the realm of graphics processing unit (GPU) computing in AI.

The main downside of TEEs in the context of PPML is that most of the available implementations are designed for use on central processing units (CPU) only. This can of course be a problem for deep learning models that have thrived because of the joint availability of large datasets and powerful GPU hardware. Some recent works attempt to address this shortcoming with proposals for GPU TEEs [81] as well as protocols leveraging a CPU TEE alongside a regular GPU while maintaining data privacy [82].

Outlook

Privacy-preserving machine learning has the potential to revolutionize AI in the medical field and beyond. It can allow researchers to avail themselves of larger datasets without infringing on the fundamental patient right to data privacy, impart a larger measure of control over personal data, and enable the provision of AI-based services under appropriate assurances of confidentiality and single-use accountability.

For the utilization of PPML to enter the mainstream, research is still required to address several crucial challenges. Federated learning systems able to train algorithms on heterogeneous data with the same predictive accuracy and performance as centralized data sharing will be required. Differential privacy research needs to address the challenging questions of granular privacy budgeting and tracking, and develop innovative solutions to apply DP to any data in the medical field such as imaging. Homomorphic encryption schemes with high accuracy and low computational overhead need to be developed, while SMPC protocols will be required to become more efficient and suffer from lower network overhead to be practical for widespread utilization.

PPML techniques evidently co-evolve with machine learning and data science. Hence, questions of human interaction with data and algorithms will have to be addressed to optimize outcomes. For example, it remains unclear how

encrypted data can be quality-controlled and curated, or how privacy guarantees interact with crucial requirements for algorithmic fairness and interpretability [83, 84]. Addressing specific legal requirements will also require targeted research. A prime example is the European General Data Protection Regulation *right to be forgotten* (Article 17), which has prompted exploration of the topic of *machine unlearning* [85]. Beyond mere privacy concerns, AI in critical applications such as health care needs to be responsible, auditable, and trustworthy. Research into verifiability of AI systems and objectively provable system attributes [86] will yield advances in this regard.

Healthcare is inherently a multi-stakeholder field. Although not specific to PPML, the questions of who owns an AI model and how they should be reimbursed are more complex in the setting of collaborative algorithm training over several sites, such as federated learning.

Nevertheless, many indicators, not least the success of the abovementioned pilot projects, suggest that PPML represents the future of AI in healthcare. The democratization of PPML tools has an important role to play in this regard. Free and open-source software libraries such as PySyft or TensorFlow Federated lower the barrier to entry into privacy-preserving AI research. Current work on software and infrastructure frameworks for healthcare-related tasks, such as the PriMIA library [87] or the German Cancer Consortium Joint Imaging Platform [88] has very recently yielded successes in end-to-end privacy-preserving deep learning on medical imaging and multi-institutional clinical data analytics, respectively. Last but not least, PPML represents a highly multifaceted field, which will require contributions by diverse teams of interdisciplinary researchers, not limited to AI and its adjacent subject but including – among others – ethics, philosophy, and jurisprudence and, perhaps most crucially, involving patients and society as a whole. OpenMined (<https://www.openmined.org>), a fully decentralized nonprofit community uniting over ten thousand individuals, aims to provide a forum for such discussion and foster interdisciplinary research in the field.

In summary, the current chapter introduced basic concepts and techniques of secure and private AI in medicine. As the field matures, it is the authors' belief that it has the potential to cause a quantum leap in nearly every area of health care, serving to unite researchers and healthcare providers from all over the world in privacy-preserving and equitable collaboration.

References

1. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafiyan H, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577(7788):89–94.
2. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med*. 2019;25(6):954–61.
3. Pinker K, Chin J, Melsaether AN, Morris EA, Moy L. Precision medicine and radiogenomics in breast cancer: new approaches toward diagnosis and treatment. *Radiology*. 2018;287(3):732–47.
4. Lu H, Arshad M, Thornton A, Avesani G, Cunnea P, Curry E, et al. A mathematical descriptor of tumor-mesoscopic-structure from computed-tomography images annotates prognostic and molecular-phenotypes of epithelial ovarian cancer. *Nat Commun*. 2019;10(1):1–11.
5. Kaassis G, Ziegelmayer S, Löhöfer F, Algül H, Eiber M, Weichert W, et al. A machine learning model for the prediction of survival and tumor subtype in pancreatic ductal adenocarcinoma from preoperative diffusion-weighted imaging. *Eur Radiol Exp*. 2019;3(1):1–9.
6. Varghese B, Chen F, Hwang D, Palmer SL, Abreu ALDC, Ukimura O, et al. Objective risk stratification of prostate cancer using machine learning and radiomics applied to multiparametric magnetic resonance images. *Sci Rep*. 2019;9(1):1–10.
7. Elshafeey N, Kotrotsou A, Hassan A, Elshafei N, Hassan I, Ahmed S, et al. Multicenter study demonstrates radiomic features derived from magnetic resonance perfusion images identify pseudoprogression in glioblastoma. *Nat Commun*. 2019;10(1):1–9.
8. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. Chexnet: radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:171105225*. (2017).
9. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomed Eng*. 2018;2(3):158.
10. Health Insurance Portability and Accountability Act; (1996). <https://www.hhs.gov/hipaa/index.html>.

11. General Data Protection Regulation; (2016). <https://eur-lex.europa.eu/eli/reg/2016/679/2016-05-04>.
12. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nature Mach Intelligence*. 2019;1(9):389–99.
13. McLennan S, Fiske A, Celi LA, Müller R, Harder J, Ritt K, et al. An embedded ethics approach for AI development. *Nature Mach Intelligence*. 2020;2(9): 488–90. <https://doi.org/10.1038/s42256-020-0214-1>.
14. Price WN, Cohen IG. Privacy in the age of medical big data. *Nat Med*. 2019;25(1):37–43.
15. Tanner A. Our bodies, our data: how companies make billions selling our medical records. Boston, MA, USA: Beacon Press; 2017.
16. Barrows RC Jr, Clayton PD. Privacy, confidentiality, and electronic medical records. *J Am Med Inform Assoc*. 1996;3(2):139–48.
17. Warren SD, Brandeis LD. The right to privacy. *Harvard Law Rev*. 1890;4(5):193–220.
18. Kaassis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat Mach Intell*. 2020;2:305–311.
19. Nissenbaum H. A contextual approach to privacy online. *Daedalus*. 2011;140(4):32–48.
20. Sweeney L. k-anonymity: a model for protecting privacy. *Int J Uncertain Fuzziness Knowl Based Syst*. 2002;10(05):557–70.
21. Bindschaedler V, Grubbs P, Cash D, Ristenpart T, Shmatikov V. The Tao of inference in privacy-protected databases. *Proc VLDB Endowment*. 2018;11(11):1715–28.
22. Narayanan A, Shmatikov V. Robust de-anonymization of large sparse datasets. In: 2008 IEEE symposium on security and privacy (sp 2008). Oakland, CA, USA: IEEE; 2008. p. 111–25.
23. Douriez M, Doraiswamy H, Freire J, Silva CT. Anonymizing NYC taxi data: Does it matter? In: 2016 IEEE international conference on data science and advanced analytics (DSAA). Piscataway: IEEE; 2016. p. 140–8.
24. Rocher L, Hendrickx JM, De Montjoye YA. Estimating the success of re-identifications in incomplete datasets using generative models. *Nat Commun*. 2019;10(1):1–9.
25. Parks CL, Monson KL. Automated facial recognition of computed tomography-derived facial images: patient privacy implications. *J Digit Imaging*. 2017;30(2):204–14.
26. Schwarz CG, Kremers WK, Therneau TM, Sharp RR, Gunter JL, Vemuri P, et al. Identification of anonymous MRI research participants with face-recognition software. *N Engl J Med*. 2019;381(17):1684–6.
27. de Sitter A, Visser M, Brouwer I, Cover K, van Schijndel R, Eijgelaar R, et al. Facing privacy in neuroimaging: removing facial features degrades performance of image analysis methods. *Eur Radiol*. 2020;30(2):1062–74.
28. Konečný J, McMahan B, Ramage D. Federated optimization: distributed optimization beyond the datacenter. *arXiv preprint arXiv:151103575*. (2015).
29. Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, Bhagoji AN, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:191204977*. (2019).
30. Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci Rep*. 2020;10(1):1–12.
31. Roth HR, Chang K, Singh P, Neumark N, Li W, Gupta V, et al. Federated learning for breast density classification: a real-world implementation. In: Domain adaptation and representation transfer, and distributed and collaborative learning. Cham: Springer; 2020. p. 181–91.
32. Ghorbani A, Zou J. Data Shapley: equitable valuation of data for machine learning. *arXiv preprint arXiv:190402868*. (2019).
33. Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the 22nd ACM SIGSAC conference on computer and communications security CCS 2015. ACM Press; 2015. <https://doi.org/10.1145/2810103.2813677>.
34. Geiping J, Bauermeister H, Dröge H, Moeller M. Inverting gradients—how easy is it to break privacy in federated learning? *arXiv preprint arXiv:200314053*. 2020.
35. Papernot N, McDaniel P, Sinha A, Wellman MP. SoK: security and privacy in machine learning. In: 2018 IEEE European symposium on security and privacy (EuroS&P). Piscataway IEEE; 2018. p. 399–414.
36. Kerkouche R, Ács G, Castelluccia C. Federated learning in adversarial settings. *arXiv preprint arXiv:201007808*. (2020).
37. EMEA. MELLODDY. Accessed: 2020-12-14. <https://www.meloddy.eu>
38. Nasirigerdeh R, Torkzadehmahani R, Matschinske J, Frisch T, List M, Späth J, et al. sPLINK: a federated, privacy-preserving tool as a robust alternative to meta-analysis in genome-wide association studies. 2020. <https://doi.org/10.1101/2020.06.05.136382>.
39. Li W, Milletari F, Xu D, Rieke N, Hancox J, Zhu W, et al. Privacy-preserving federated brain tumour segmentation. In: International workshop on machine learning in medical imaging. Shenzhen, China: Machine Learning in Medical Imaging; 2019. p. 133–41.
40. Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. *npj Digit Med*. 2020;3(1) <https://doi.org/10.1038/s41746-020-00323-1>.
41. Vaid A, Jaladanki SK, Xu J, Teng S, Kumar A, Lee S, et al. Federated learning of electronic health records improves mortality prediction in patients hospitalized with COVID-19. 2020. <https://doi.org/10.1101/2020.08.11.20172809>.

42. Dwork C. Differential privacy: a survey of results. In: International conference on theory and applications of models of computation. Xi'an, China: TAMC; 2008. p. 1–19.
43. Dwork C, Roth A, et al. The algorithmic foundations of differential privacy. *Found Trends Theoret Comp Sci*. 2014;9(3–4):211–407.
44. Mironov I. Rényi differential privacy. In: 2017 IEEE 30th computer security foundations symposium (CSF). Piscataway: IEEE; 2017. p. 263–75.
45. Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. In: Theory of cryptography conference. Berlin: Springer; 2006. p. 265–84.
46. Miresghallah F, Taram M, Jalali A, Elthakeb AT, Tullsen D, Esmaeilzadeh H. A principled approach to learning stochastic representations for privacy in deep neural inference. arXiv preprint arXiv:200312154. (2020).
47. Jordon J, Yoon J, van der Schaar M. PATE-GAN: generating synthetic data with differential privacy guarantees. In: International conference on learning representations. La Jolla, CA, USA: Published by the International Conference on Representation Learning; 2018.
48. Papernot N, Abadi M, Ásif Erlingsson, Goodfellow I, Talwar K. Semi-supervised knowledge transfer for deep learning from private training data; (2017).
49. Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, et al. Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. New York, NY, United States: Association for Computing Machinery; 2016. p. 308–318.
50. Francis P, Probst-Eide S, Obrok P, Berneanu C, Juric S, Munz R. Diffix-Birch: extending DiffixAspen; (2019).
51. Kukačka J, Golkov V, Cremers D. Regularization for deep learning: a taxonomy. arXiv preprint arXiv:171010686. (2017).
52. Kim JW, Jang B, Yoo H. Privacy-preserving aggregation of personal health data streams. *PLoS One*. 2018;13(11):e0207639. <https://doi.org/10.1371/journal.pone.0207639>.
53. Azencott CA. Machine learning and genomics: precision medicine versus patient privacy. *Philos Trans R Soc A Math Phys Eng Sci*. 2018;376(2128):20170350. <https://doi.org/10.1098/rsta.2017.0350>.
54. Berger B, Cho H. Emerging technologies towards enhancing privacy in genomic data sharing. *Genome Biol*. 2019;20(1) <https://doi.org/10.1186/s13059-019-1741-0>.
55. Jiang Y, Wang C, Wu Z, Du X, Wang S. Privacy-preserving biomedical data dissemination via a hybrid approach. *AMIA Ann Symp Proc*. 2018;2018:1176. American Medical Informatics Association
56. Winslett M, Yang Y, Zhang Z. Demonstration of damson: differential privacy for analysis of large data. In: 2012 IEEE 18th international conference on parallel and distributed systems. Piscataway: IEEE; 2012. p. 840–4.
57. Choudhury O, Gkoulalas-Divanis A, Salonidis T, Sylla I, Park Y, Hsu G, et al. Differential privacy-enabled federated learning for sensitive health data. arXiv preprint arXiv:191002578. (2019).
58. Dankar FK, El Emam K. Practicing differential privacy in health care: a review. *Trans Data Priv*. 2013;6(1):35–67.
59. Yang M, Lyu L, Zhao J, Zhu T, Lam KY. Local differential privacy and its applications: a comprehensive survey. arXiv preprint arXiv:200803686. (2020).
60. Xiong X, Liu S, Li D, Cai Z, Niu X. A comprehensive survey on local differential privacy. *Secur Commun Netw*. 2020;2020:1–29. <https://doi.org/10.1155/2020/8829523>.
61. Hesamifard E, Takabi H, Ghasemi M. Cryptodl: deep neural networks over encrypted data. arXiv preprint arXiv:171105189. (2017).
62. Gilad-Bachrach R, Dowlin N, Laine K, Lauter K, Naehrig M, Wernsing J. Cryptonets: applying neural networks to encrypted data with high throughput and accuracy. In: International conference on machine learning. New York, NY, United States: Association for Computing Machinery; 2016. p. 201–10.
63. Cheon JH, Kim A, Kim M, Song Y. Homomorphic encryption for arithmetic of approximate numbers. In: Takagi T, Peyrin T, editors. *Advances in cryptology – ASIACRYPT 2017*. Cham: Springer International Publishing; 2017. p. 409–37.
64. Juvekar C, Vaikuntanathan V, Chandrasekaran A. GAZELLE: a low latency framework for secure neural network inference. In: Proceedings of the 27th USENIX conference on security symposium. Baltimore, MD, USA: SEC’18. USENIX Association; 2018. p. 1651–68.
65. Chillotti I, Gama N, Georgieva M, Izabachène M. TFHE: fast fully homomorphic encryption over the torus. *J Cryptol*. 2020;33(1):34–91.
66. Blatt M, Gusev A, Polyakov Y, Goldwasser S. Secure large-scale genome-wide association studies using homomorphic encryption. *Proc Natl Acad Sci*. 2020;117(21):11608–13. <https://doi.org/10.1073/pnas.1918257117>.
67. Guo J, Liu Z, Lam KY, Zhao J, Chen Y, Xing C. Secure weighted aggregation in federated learning; (2020).
68. Acar A, Aksu H, Uluagac AS, Conti M. A survey on homomorphic encryption schemes: theory and implementation. *ACM Comput Surveys (CSUR)*. 2018;51(4):1–35.
69. Zhao C, Zhao S, Zhao M, Chen Z, Gao CZ, Li H, et al. Secure multi-party computation: theory, practice and applications. *Inf Sci*. 2019;476:357–72.
70. Evans D, Kolesnikov V, Rosulek M. A pragmatic introduction to secure multi-party computation. *Foundat Trends® Privacy Secur*. 2017;2(2–3):7ff.
71. Shamir A. How to share a secret. *Commun ACM*. 1979;22(11):612–3.

72. Ryffel T, Pointcheval D, Bach F. ARIANN: low-interaction privacy-preserving deep learning via function secret sharing. arXiv preprint arXiv:200604593. (2020).
73. Marwan M, Kartit A, Ouahmane H. Applying secure multi-party computation to improve collaboration in healthcare cloud. In: 2016 third international conference on systems of collaboration (SysCo); (2016). p. 1–6.
74. Tso R, Alelaiwi A, Rahman SMM, Wu ME, Hossain MS. Privacy-preserving data communication through secure multi-party computation in healthcare sensor cloud. *J Signal Process Syst.* 2016;89(1):51–9. <https://doi.org/10.1007/s11265-016-1198-2>.
75. Veeningen M, Chatterjee S, Horváth AZ, Spindler G, Boersma E, van der SPEK P, et al. Enabling analytics on sensitive medical data with secure multi-party computation. In: MIE. Switzerland: European Federation of Medical Informatics, Le Mont-sur-Lausanne; 2018. p. 76–80.
76. Reichert L, Brack S, Scheuermann B. Privacy-preserving contact tracing of COVID-19 patients. IACR Cryptol ePrint Arch. 2020;2020:375.
77. Sabt M, Achemlal M, Bouabdallah A. Trusted execution environment: what it is, and what it is not. In: 2015 IEEE Trustcom/BigDataSE/ISPA, vol. 1. Los Alamitos: IEEE; 2015. p. 57–64.
78. Haralampieva V, Rueckert D, Passerat-Palmbach J. A systematic comparison of encrypted machine learning solutions for image classification. Proceedings of the 2020 workshop on privacy-preserving machine learning in practice. 2020;ISBN: 9781450380881 Publisher: ACM. <https://doi.org/10.1145/3411501.3419432>.
79. Tsai CC, Porter DE, Vij M. Graphene-sgx: A practical library {OS} for unmodified applications on {SGX}. In: 2017 {USENIX} annual technical conference ({USENIX} {ATC} 17); 2017. p. 645–658.
80. Lindell Y. The security of Intel SGX for key protection and data privacy applications. 2018; p. 13. Available from: <https://www.unboundtech.com/wp-content/uploads/2020/09/security-of-intelsgx-key-protection-data-privacy-apps.pdf>.
81. Volos S, Vaswani K, Bruno R. Graviton: trusted execution environments on GPUs. In: 13th USENIX symposium on operating systems design and implementation (OSDI 18). Carlsbad: USENIX Association; 2018. p. 681–96. Available from: <https://www.usenix.org/conference/osdi18/presentation/volos>.
82. Tramer F, Boneh D. Slalom: fast, verifiable and private execution of neural networks in trusted hardware. arXiv:180603287 [cs, stat]. 2018;ArXiv: 1806.03287. Available from: <http://arxiv.org/abs/1806.03287>.
83. Harder F, Bauer M, Park M. Interpretable and differentially private predictions. AAAI. 2020;34:4083–90.
84. Agarwal S. Trade-offs between fairness, interpretability, and privacy in machine learning. UWSpace; (2020). Available from: <http://hdl.handle.net/10012/15861>.
85. Bourtoule L, Chandrasekaran V, Choquette-Choo C, Jia H, Travers A, Zhang B, et al. Machine unlearning. arXiv preprint arXiv:191203817. (2019).
86. Brundage M, Avin S, Wang J, Belfield H, Krueger G, Hadfield G, et al. Toward trustworthy AI development: mechanisms for supporting verifiable claims. arXiv preprint arXiv:200407213. (2020).
87. Kaassis G, Ziller A, Passerat-Palmbach J, et al. End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nat Mach Intell.* 2021;3: 473–484. <https://doi.org/10.1038/s42256-021-00337-8>.
88. Scherer J, Nolden M, Kleesiek J, Metzger J, Kades K, Schneider V, et al. Joint imaging platform for federated clinical data analytics. *JCO Clin Cancer Inform.* 2020;4:1027–38. <https://doi.org/10.1200/cci.20.00045>.



Artificial Intelligence for Medical Decisions

11

Albert Buchard and Jonathan G. Richens

Contents

Introduction	160
Automation of Decision-Making in Healthcare	160
Taxonomy of Medical Decisions	161
Logic-Based Methods	162
The Language of Logic	162
Knowledge Representation and Reasoning	163
Beyond First-Order Logic	163
Learning from Data	164
Statistical Modelling and Machine Learning	164
Three Machine-Learning Approaches	165
The Impact of the Deep Learning Revolution	167
Combinatorial Optimization Methods	168
Reinforcement Learning for Sequential Decision-Making	168
Bayesian Models for Decision Support	169
Bayesian Networks for CDSS	170
The Need for Causality in Clinical Decision-Making	171
Explainability, Interpretability, and Fairness	173
Conclusion	174
References	174

Abstract

While artificial intelligence has been promised to revolutionize healthcare for decades, the field has yet to reap the benefits of automation compared to other domains. This chapter presents an overview of the vast field of decision-making in healthcare and proposes a functional classification of decision tasks. We aim to equip the reader with a working knowledge of

A. Buchard
Service de Psychiatrie adulte, Hopitaux Universitaires de Genve, Geneva, CH, Switzerland
e-mail: albert.buchard@hcuge.ch

J. G. Richens (✉)
AI research, Babylon Health, London, UK
e-mail: jonathan.richens@babylonhealth.com

methods for clinical decision-making, including logic-based, learning-based, and Bayesian methods. We then introduce the reader to modern causal approaches to decision-making and reiterate the need for externally validated, explainable, and fair methods. Beyond the individual strengths and weaknesses of these methods, we argue that future AI products should embrace hybrid approaches and focus not only on improving clinical outcomes but also on offering a streamlined design that accelerates processes and facilitates clinical practice.

Keywords

Clinical decision-making · Artificial intelligence · Decision-making · Automated decision-making · Causality

Introduction

Automation of Decision-Making in Healthcare

Since Antiquity, Medical practice has been organized around codified rules of conduct such as the Code of Hammurabi (1740 BC) or the Hippocratic oath (circa BC 460–360), derived from an understanding of physiopathology as well as political and religious constraints [1]. However, it was not until the development of modern decision theory in the 1950s [2, 3] that medical decision-making found a mathematical basis in symbolic logic and probability [4]. From then and until the early 1970s, the first wave of research focused on statistical approaches to perform diagnosis [5]. However, these methods lost their appeal when the field of AI emerged, built around knowledge representation and logic [6]. The 1980s gave rise to many logic-based “*Expert-Systems*” which, to this day, are among the most successful and finalized clinical decision support systems (CDSS) [7, 8].

In the 1990s and 2000s, during and after the *second AI winter*, research and institutional interest in AI and automated medical decision-making became more sporadic, and except for a few

projects [9, 10], the development of expert-systems dwindled. However, during that time, the internet facilitated the dissemination of medical knowledge, and computer systems became prevalent across most healthcare domains, especially for administration, data management, data analysis, and medical imaging. The use of computer systems in clinical decision support was relegated mainly to hospital EHR subsystems, such as drug dosage and interaction tools, or for simple automated detection, such as rule-based alarms for continuous monitoring in critical care units [11] or ECG machines able to perform diagnosis [12].

Besides the natural reticence and prudence of the medical culture [13], the slow development of automated decision tools not only highlights the complexity of decision-making in healthcare settings and the difficulty to improve clinical outcomes, but also the challenge of implementing AI research into well-designed products. The feeling that the science has not yet been made into acceptable technology is not new [6, 14], and it can be surprising that after 70 years of research and development, hospitals and clinician are still not equipped with computerized assistants. However, with recent innovations in Machine Learning, the last decade has seen a renewed interest in AI in the private and public sector [15]. In tandem with these theoretical advances, numerous innovations in basic science in the past decade have allowed us to gather new and complex types of data (e.g., fast genome sequencing and gene editing [16]). Software and hardware development has also gone through a tremendous revolution, including the creation of many new languages, interactive development environments, and programming frameworks to formalize and speed up development. The CUDA framework, for example, developed by NVIDIA, facilitated access to the parallel processing of Graphical Processing Units (GPUs) for non-graphics tasks, which later helped improve Machine Learning performance [17]. The proliferation of cloud storage and cloud computing has enabled large-scale deployment and maintenance of complex projects. Many of the digital technologies of the ‘90s and 2000s have matured, accompanied by a solid product design know-how reflected in the importance given to new roles

such as UX designers. As a result, since 2010, the market for patient-centric medical apps and smart-wear has bloomed [18]. Overall, most of these apps were developed outside of the healthcare sector, had poor retention rate [19] and – due to gaps in the regulatory frameworks – were improperly evaluated [20]. Recently, the mobile health sector has started a transition towards better frameworks of evaluation, publishing academic papers, and evaluating clinical outcomes using Randomized Control Trials [21, 22]. We are also observing an increasing interest in telemedicine solutions allowing direct patient access to their medical files, remote monitoring and consultations [23].

These recent innovations highlight the maturing interface between the medical and the technology sectors. They are strong indicators that the generation of medical products of the 2020s could finally offer the performance, automated functionality, ease of integration, and design quality necessary to deliver on the promises of AI in clinical decision-making – kick-starting the long-awaited revolution in healthcare access and practices.

Taxonomy of Medical Decisions

Healthcare systems are notoriously complex and comprise thousands of agents making decisions individually or collectively over multiple scales, from patients and clinicians to policymakers. Before introducing the AI methods used to solve complex decision-making tasks in healthcare, we must sketch the boundaries of these tasks and offer a general taxonomy inspired in part from previous works [24, 25]. The majority of the decision-making tasks in healthcare can, in principle, be automated or supported by computational methods. We can distinguish five decision domains (see Fig. 1). Across those domains, automated methods hold the promise of reducing human error, enabling better-individualized decisions, reducing costs, improving access to healthcare, and accelerating research and innovation.

Firstly, automated methods can be applied to basic and medical research to understand better human physiology and physio-pathology of diseases using large scale datasets or to discover new treatments. The RX project, developed by Blum in

CLINICAL TASKS		BASIC SCIENCE AND MEDICAL RESEARCH		LOGISTICS		HEALTHCARE POLICY	
Diagnosis	To discover the cause of a patient's abnormal condition	Statistical Modelling And Simulation	Model generative processes to test hypotheses or forecast a system's evolution (e.g. agent based modelling)	Resource Allocation	Workforce management tasks (e.g. nurse to patient assignment problem, operating room scheduling), patient planning (e.g. inpatients planning or geographical patient allocation), crisis management, but also other sub-tasks in more significant planning efforts such as hospital department design	Evidence Synthesis	Tools for automated or semi-automated analysis and summarization of research data and expert knowledge in order to facilitate decision making
- Passive Diagnosis	Single step diagnosis process which only rely on currently available information	Experimental Data Interpretation	Treat, visualise and analyse complex data (e.g. 3D reconstruction and labelling of microscopy data)	Quality Assessment	Tasks aimed at facilitating or performing automated assessments of processes in place in a healthcare setting. Such as evaluating the correct use and efficiency of decision algorithms and clinical pathways inside a hospital	Clinical Guideline Definition	Produce decision algorithms to homogenise and improve clinical practices
- Active Diagnosis	To plan an efficient data gathering sequence, and update it as new information is available, in order to make an accurate final diagnosis	Disease Understanding	Discover new conditions or partition existing conditions in a clinically or physiologically relevant manner (e.g. gene network cluster analysis)	Other support tasks	Coordinate or enhance productivity and communication across an organisation, or perform administrative tasks such as generating and processing financial statements	National And Supranational Policy	Define large scale health policies, financed by taxes for disease prevention, such as cancer screening, food and drug regulation, and publicly financed community care programs.
Prognosis	To forecast the likely course of a medical situation	Drug Discovery	Generate and evaluate new drug candidates (e.g. quantitative structure-activity relationship models)				
Therapeutic decision	To devise one or several plans to remedy, or prevent, a patient's condition.						
- Single-stage	To make a single therapeutic decision						
- Multi-stage	To make several therapeutic decision in time, adapting to the patient's response						
Surveillance	To implement a plan to monitor the evolution of a patient's condition during treatment, as well as to detect adverse effects and complications						
Physical Procedures	Embodyed tasks which require physical interactions with the patient, such as performing a clinical examination, an invasive test, or surgery						
Medical Imaging	Support tasks related to the treatment and analysis of medical imaging data						
- Enhancement / Reconstruction	Tasks aimed at improving the quality of medical imaging such as denoising, upscaling of low resolution imaging, or volume reconstruction						
- Detection / Segmentation	Detect, label, and/or contour anomalies or anatomical structures						
- Registration	Map one or several imaging methods to a single reference frame (e.g. registration of MRI and stereotactic radiography for deep brain stimulation implantations)						
Other support tasks	Many other tasks asked of physician during their practice can be automated. For example, speech processing, note taking and summarizing, translation, drug interaction detection, administrative tasks (insurance, certificates), or therapeutic education						

Fig. 1 Decision-making tasks in healthcare span across four broad categories: basic science, clinical, logistics, and policy-making. We present a non-exhaustive taxonomy of the main tasks in each category

1982, was one of the first such systems used to automate hypothesis generation and testing from a knowledge base of facts and statistical knowledge [26]. A more recent example is in drug discovery, where Deep Learning methods have been used to develop better Quantitative Structure-Activity relationship models and improve the selection process for promising molecular candidates [27]. Other examples include systems able to match patients to available clinical trials [28], predict protein folding [29], build disease progression models [30], or reconstruct 3D volumes from large microscopy datasets [31].

The second class of decisions relates to clinical care. This category covers most of the daily decision-making of medical doctors: triage, diagnosis, prognosis, treatment selection, but also all the support tasks surrounding these. These functional categories conceal a rich diversity, such as the nature and precision of the expected output, or the time scales at which decisions should be made. For example, evaluating the risk of lifestyle decisions over the lifespan or predicting the immediate evolution of a hemorrhagic choc are both prognostic tasks. Similarly, in the case of a cancer patient, treatment decisions can be deliberate, multi-disciplinary, and thorough, while life-threatening situations may require fast, purposeful, and sometimes approximate decisions [32].

The third domain relates to logistics and planning, where automated methods could help to better distribute resources, or to test and improve internal processes. For example, computational approaches could help forecast MRI scanners utilization in a hospital, better distribute patients across hospitals in a city [33], or help in the distribution of life-saving resources in a crisis [34].

Finally, computational methods are already used to support decisions regarding healthcare policies. This type of decision-making process is complex and multi-disciplinary and relies on data-driven insights, experts' opinions, as well as economic and political constraints. Statistical approaches are already vital to developing and refining clinical guidelines, as well as for deciding on global health policies at the national and international level [35]. The recent innovations in data-

driven methods, the increased access to larger datasets and the structuring of medical and research knowledge in large knowledge bases provide new opportunities for making sense of the vast amount of information now available to experts [36].

While the logic-based, learning-based, combinatorial optimization and Bayesian methods used to solve these tasks are theoretically distinct, the majority of clinical decision support systems (CDSS) are hybrid systems [24, 37]. To equip the reader with a working knowledge of AI methods, we present the general ideas of each of these approaches in the following sections.

Logic-Based Methods

Clinical knowledge is highly structured, both in terms of medical concept hierarchies and through the guidelines and clinical pathways which steer clinical decisions and draw the boundaries of malpractice. Logic-based methods can naturally encode these structures, making them usually safe and interpretable [24]. As a result, they are some of the most trusted and successfully deployed AI solutions for automated decision-making in medicine with projects such as MYCIN [7], INTERNIST-I [38], CADUCEUS [8], which employed logic-based reasoning solely or in part. This approach is often referred to as Knowledge-based AI because inferences are generated by applying logical rules to stored knowledge, usually gathered from experts.

The Language of Logic

Logic-based systems generate decisions by representing truth statements such as 'the patient has a fever' symbolically and combining these into logical propositions whose truth value generates the desired output. Depending on the logical formalism used, these propositions may include different types of objects. Propositional logic, or 0th-order logic, only deals with manipulations of atomic statements such as:

$$\begin{array}{l} \text{patient_has_fever} \Rightarrow \\ \text{administer_paracetamol. } p \Rightarrow q \end{array} \quad (1)$$

which represents a treatment rule whereby paracetamol is administered if and only if the patient has a fever.

On the other hand, higher-order logic, such as first-order logic, can handle universal quantifiers and variables, as well as predicates and functions over these. In turn, it allows for compositionality and reusable structures for constructing propositions. For example, instead of building two statements for p : “Core temperature of the patient is above 37.5 °C” and q : “Core temperature of the patient is above 39 °C,” with p and q stored and assigned through specific code, first-order logic enables one to express those two statements as $Gt(t(x), 37.5)$ and $Gt(t(x), 39)$ – a composite of the functions $t(x)$, which returns the core temperature of x , and the greater-than function $Gt(x, y)$ which is True if $x > y$. Variables can also be quantified as in:

$$\forall x(\text{Patient}(x) \wedge Gt(t(x), 38)) \Rightarrow \text{Fever}(x), \quad (2)$$

which corresponds to the sentence *For all patients x , if x has a core temperature above 38 then x has a fever*. While such composite statements would also need to be stored, their implementation is now modularized and offers many advantages. For example, it allows experts to work with composable objects during knowledge elicitation, which can be integrated into a readable user interface. It also better reflects how the code is implemented, with reusable and composable functions (i.e., accessing core temperature sensor) working on known objects (i.e., the patient).

Knowledge Representation and Reasoning

In real-world applications, the set of propositions known to the system is usually stored in a knowledge base (KB) and represented differently depending on the use-case. In “Rules-based” systems the KB emulates human

expertise and reasoning processes. It is comprised of sequences of If-Then conditions in a branching fashion, effectively forming decision trees of the form “*If P is True then do A , else do B* ,” with A and B being either a decision or another condition. Rule-based systems, such as MYCIN [7], are usually purely deductive (*forward-chaining*), and generate decisions as the tree is explored (see Fig. 2).

Other systems use an *ontological KB*, which defines concepts by their relationships to each other and answers logical queries via a combination of deduction, abduction and heuristics. Ontological KBs represent knowledge in the form “*A Relation B*,” where A and B are concepts, and *Relation* denotes a primitive relationship between concepts. For example, “Meningitis IS AN Infectious Disease” uses the relation “IS AN” to encode Infectious Disease as a parent concept of Meningitis. Queries of the form “What is Meningitis?” can be answered deductively by a lookup for all concepts X which satisfies the proposition “*Meningitis IS A X*,” for example returning {“Infectious Disease,” “Disease,” “Medical Finding”}. Examples of abductive queries include “What is the most likely cause of this patient’s fever?” which requires a ranking of compatible causes (“ X CAUSES Fever”) via a heuristic scoring function. For example, in INTERNIST-1 [38], characteristic disease profiles, known as *frames*, are used to match a patient clinical presentation to the “closest” known profile.

Beyond First-Order Logic

Despite their large adoption, logic-based approaches are unable to deal with high combinatorial complexity. For example, with a hundred symptoms, the number of valid symptoms combinations and potential rules easily exceeds the number of atoms in the observable universe. Another drawback is their inherent weakness in handling uncertainty and ambiguity, which are found throughout medical decision-making. To better handle the stochastic nature of real-world tasks, a new family of methods based on Fuzzy-logic was developed. Fuzzy-Logic arose from the

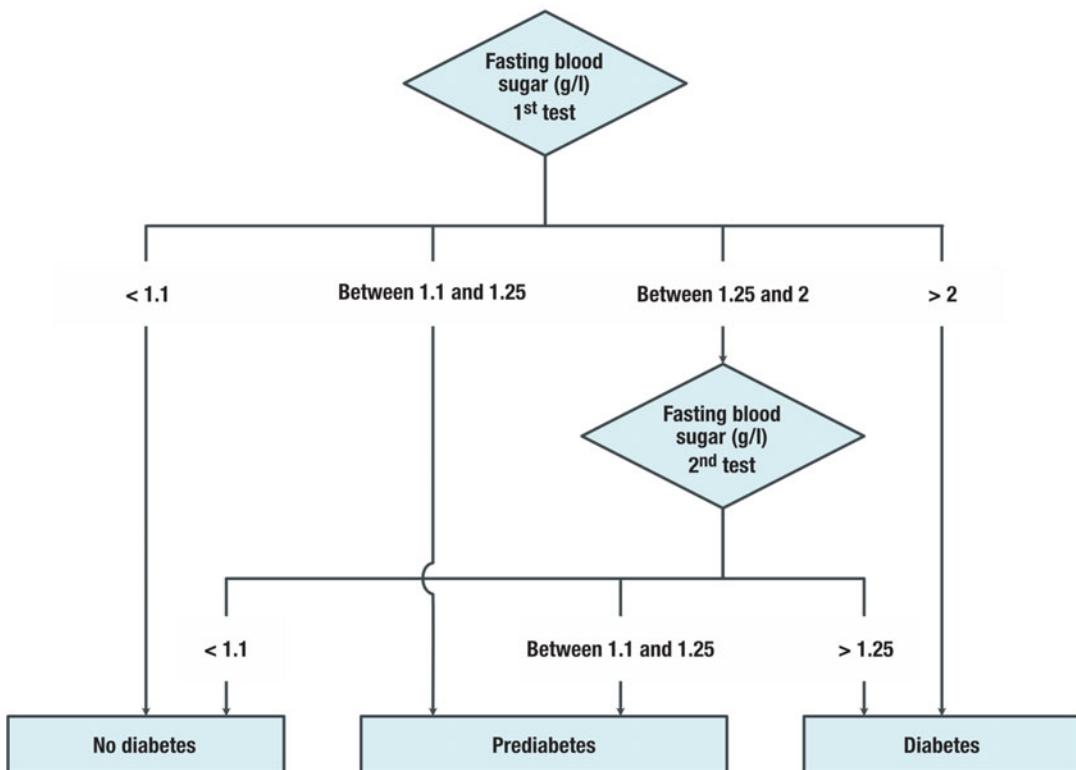


Fig. 2 Example of a decision tree used for diagnosing diabetes

concept of Fuzzy sets [39], and deals with approximate truth values by relaxing the notion of set membership to account for partial membership. While it is not meant to produce valid probability distributions, Fuzzy logic allows for incorporating scoring functions to weight propositions and rank or combine the resulting decisions. This extension to logic-based methods can deal with ambiguous situations and facilitates knowledge elicitation from experts in areas where medical knowledge is imperfect [24]. Fuzzy logic CDSS are still being actively researched [40], for example, in radiotherapy treatment planning [41], or management of urinary tract infections [42].

The methods presented in the next sections resolve some of the issues of logic-based systems and their combination with logic-based approaches (e.g., Markov Logic Networks) can offer best-of-both-world solutions as a “judicious combination of categorical and probabilistic reasoning” [37].

Learning from Data

Statistical Modelling and Machine Learning

Since the 1960s, statistical modelling has transformed the way we conduct clinical research [43], has shaped the common language of Evidence-Based Medicine [44], and defined the quality standards for experiments and data analysis. It could be surprising to realize that those applied statistics methods, well known in the medical field, are in fact Machine Learning (ML) methods, in that they construct models by *fitting* them to data, optimizing for the best parameters in order to minimize prediction error. This defining characteristic means that, unlike knowledge-based methods, ML methods do not rely on experts to craft decision rules or select model parameters. While the models used in statistical analysis are

typically simple and transparent, ML models can be highly complex, in some cases comprising of billions of parameters [45]. While simple statistical models are used to solve *inference* tasks, e.g., answering queries such as “How much does event A increase the risk of disease B?” ML methods have excelled at *prediction*, which aim to approximate unseen data points [46]. In the following section, we outline the main principles and medical applications of statistical machine learning, and further explore the trade-off between the predictive power of ML methods versus their ability to answer direct associative queries (see [7]).

Three Machine-Learning Approaches

Learning can be defined as the process of discovering and storing functional knowledge about invariant properties of the world, patterns conserved in time, in order to make predictions. Machine learning methods are commonly divided into three main classes, corresponding to three different types of invariant property and task

objectives: *Supervised Learning*, *Unsupervised Learning*, and *Reinforcement Learning* (see section “[Reinforcement Learning for Sequential Decision-Making](#)”).

In *Supervised Learning* the task is to find the best parameters of a model able to map a set of predictors X into a target variable Y (see Fig. 3). This class of problems is ubiquitous in applied statistics. In biomedical research, for example, one would try to estimate the impact of a specific treatment (X) on survival (Y) using a linear model (see Fig. 4).

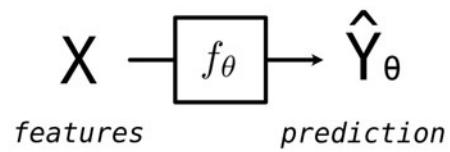
The arsenal of methods applicable to supervised learning tasks is vast, among them simpler approaches like Classification and Regression Trees (CART) [47] and Generalized Linear Models (e.g., Logistic Regression) [48] are frequently used in the medical literature. Efforts were made in recent years to facilitate the adoption of more complex approaches by non-expert and clinicians [49]. Nevertheless, models such as Structural Equation Models [50], Support Vector Machines [51], or Neural Networks [52], still necessitate in-depth knowledge in specific tools

A.

X_1	Y_1
X_2	Y_2
:	
X_n	Y_n

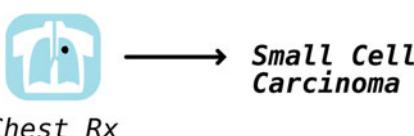
features *target*

B.



$$\theta^* = \underset{\theta}{\operatorname{argmin}} \mathcal{L}(Y, \hat{Y}_{\theta})$$

C.



D.

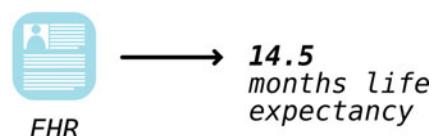


Fig. 3 In Supervised Learning, the dataset is composed of two types of variables: the features (X) from which the model is to predict one or several target variables (Y). **B.** The task is to find the optimal parameters θ^* of a model f_{θ} so that the estimate Y^{θ^*} is as close as possible to the known target Y , which corresponds to minimizing the loss function (Y, Y^{θ}) . The nature of defines the task objective and what the parameters will be optimised to solve. For example, additional constraints may be

added to the estimation error of Y to promote specific properties of the model, e.g. adding an L1 regularisation term to promote sparsity (LASSO). **C. and D.** Depending on the nature of the target variable Y , we classically separate supervised learning tasks into two subtypes: when Y is a discrete variable (e.g. tumour label, C.), the task is called *classification*, and when Y is a continuous (e.g. life expectancy, D.), the task is called *regression*.

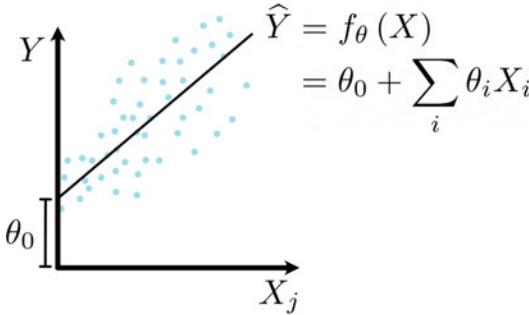


Fig. 4 f_θ is a model with parameters θ , linear in the input variables $X = \{X_0, \dots, X_k\}$. A linear model's simplicity allows estimating the magnitude of a direct and linear association between any predictor X_j and the target variable Y , expressed by a single parameter θ_j .

ad programming languages, which limits their diffusion into epidemiology and clinical research.

The second family of ML approaches is *Unsupervised Learning*, so-called because learning occurs in the absence of “supervision signal” contained in the target labels Y . Instead, the objective is to find a compressed and informative representation Z of the data X , which can explain the data or from which we could reconstruct the data using a *decoder* function (see Fig. 5). The relative importance given to the representation and the decoder depends on the specific method, with certain methods only interested in obtaining a useful representation Z (e.g. clustering), others in learning an efficient decoder to generate X (e.g., GAN, Autoregressive models). The encoder function, which transforms X into Z , and the decoder function, which recovers \hat{X} from Z , can be learned or be known a priori. One simple class of unsupervised learning algorithms is the binning algorithms used to build histograms. Indeed, a histogram is a compressed representation of continuous data, requiring only one number per bin. Using a sampling scheme as decoder, a dataset can then be reconstructed by sampling from the distribution described by the histogram. Another classic example is the task of data compression, where the goal is to find the encoding scheme which yields the smallest description (Z) while being able to generate a faithful reconstruction of the original data (\hat{X}) (see [53]).

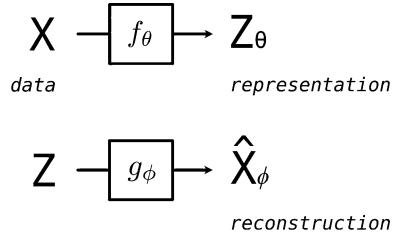


Fig. 5 In unsupervised learning, the objective is to find a representation Z of the data X . The *latent* representation Z is not directly measured but can be used to visualize, explain or reconstruct the data. Here, the invariant properties are captured both in the *encoder* f_θ , a function with parameters θ which maps X to Z_θ , as well as the *decoder* g_ϕ which is able to reconstruct \hat{X}_ϕ from Z

Unsupervised learning methods routinely used in applied statistics include clustering approaches (e.g., k-means clustering), and dimensionality reduction methods such as Factor Analysis, Principal Component Analysis, or Singular Value Decomposition. Although, in recent years the Deep Learning revolution has allowed to learn more complex encoder and decoder functions, the Deep Auto-Encoder architecture being one of the first examples [54], followed by Variational Auto-Encoders [55], Adversarial Approaches (e.g., GAN [56]), Autoregressive models [57], Normalizing Flows [58], and more recently self-supervised approaches relying on Contrastive Loss [59].

Along with advances in optimization methods, the machine learning toolbox has grown steadily since the 1960s. For example, the 1986s algorithm *ID3* allowed to learn decision trees [60], and Support Vector Machines (SVMs) [61] were able to learn decision boundaries in large input spaces. The idea that ML would become an essential tool for automated decision-making in healthcare is certainly not new [26]. However, before the turn of the century, these learning methods found limited applications in the medical field and were not deployed at scale. It started to change in the last two decades due, in part, to the digitization and storage of large medical data sets [62], as well as the increased computational power and scalability of modern algorithms (e.g., the kernel trick in SVMs). Recently, the field of AI, its reach, application, and most of all, its funding was profoundly

transformed by the development of modern Deep Learning methods.

The Impact of the Deep Learning Revolution

The 2010s have seen the rise of a transformative class of AI methods called Deep Learning (DL), which rely on large Neural Networks (NNs) models [63]. The architecture of NNs is inspired by the nervous system and was first introduced in Rosenblatt's seminal 1958 paper on *Perceptrons* [64] in an effort to describe a quantitative

approach compatible with the Hebbian Theory of synaptic plasticity. Those networks are organized hierarchically in layers composed of computational units called “*neurons*,” from the input layer to the output layer, with in between a series of “*hidden*” layers. Each unit is optimized to represent a simple function of its input, unique to each neuron, which is the composition of a linear function and a nonlinear *activation function* (e.g., sigmoid or ReLU [65]). For each neuron, the parameters of its linear function are optimized (i.e., “learned”) so that its overall participation to the error is minimized (see Fig. 6). Under certain assumptions, NNs can provably approximate any

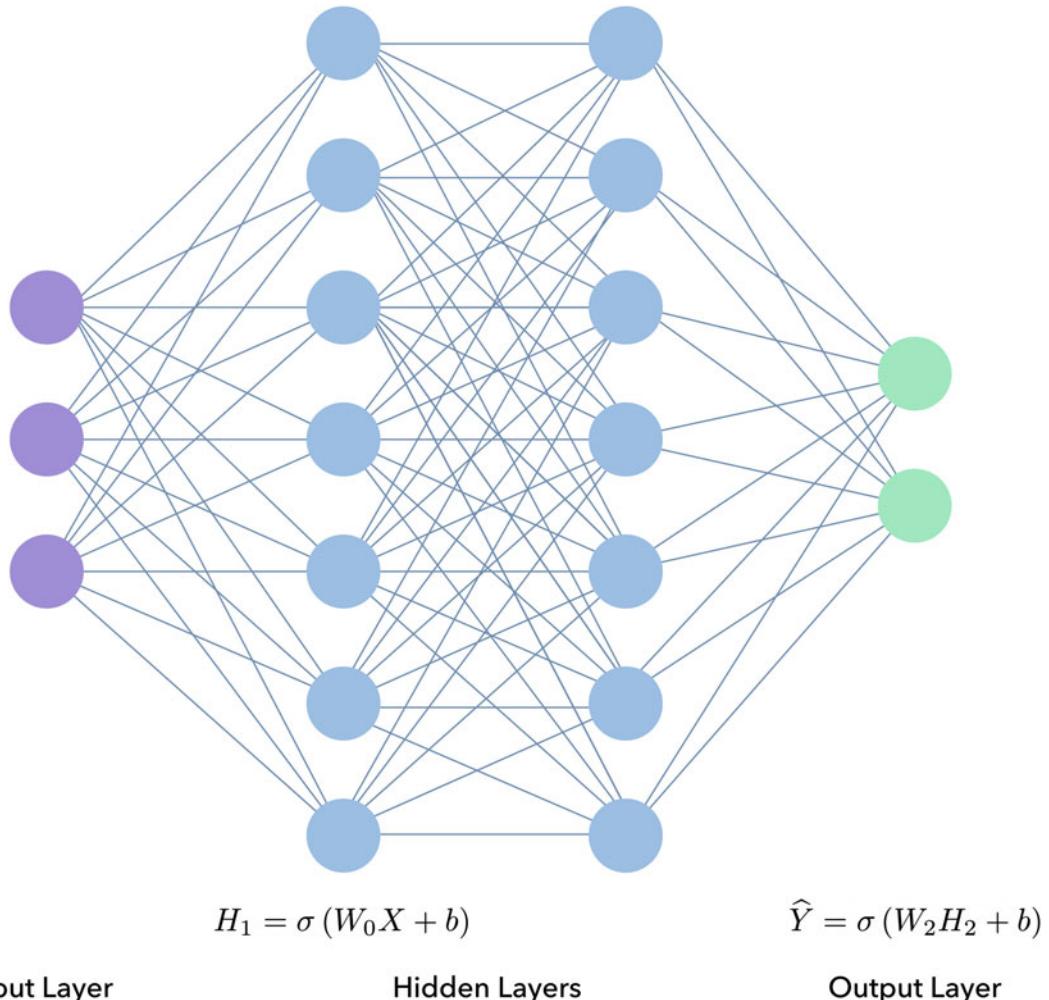


Fig. 6 Example of a feed-forward neural network taking $X \in \mathbb{R}^3$ as input, with W_k the weight matrix for a layer k , and σ the activation function (e.g. sigmoid or ReLU). The output \hat{Y} is a vector in \mathbb{R}^2

function with arbitrarily small error [66]. They are also distributed algorithms, with each neuron treating the information independently of other neurons on the same layer. Hence, same layer operations can be computed in parallel and boil down to large matrix multiplications well suited for increasingly powerful Graphical Processing Units (GPUs), which considerably speeds up learning and prediction.

These characteristics, along with several key innovations in the training methods and architectures, have enabled neural networks to efficiently find patterns inside complex, messy, and unstructured datasets beyond the limits of human perception and processing abilities. It explains the success of DL in the field of medical diagnosis from medical imagery, which will be largely covered in the next chapters, in particular in its application both as a tool to help process images and isolate critical regions, as well as a stand-alone automated diagnosis tool. DL's ability to exploit highly complex patterns could revolutionize our ability to classify, diagnose, and treat diseases earlier and with more precision [67]. However, despite its successes, DL methods typically lack the explainability and safety requirements that are central to medical practice – an issue that we explore in section “[Explainability, Interpretability, and Fairness](#).”

Combinatorial Optimization Methods

Combinatorial optimization methods were devised to speed up the search for valid solutions in well-defined systems that require solving for several interdependent variables. Rather than exploring all potential solutions blindly (see brute-force

search Fig. 7), they rely on the structure of the problems to improve efficiency. They play a crucial role in automation and are a central part of the AI toolbox. The dependencies between decisions can be causal, in time, due to the sequential nature of the problem. For example, when playing chess or when looking for the shortest path between two points. The dependency could also be through shared constraints, for example, a limited resource, such as when looking for an optimal nurse-to-patient assignment.

While some may not consider those methods as AI, it remains that it is such variations on the classic search algorithms that defeated Kasparov at chess in 1996 (alpha-beta pruning [68]), and more recently won against Lee Sedol, the world champion in the game of Go (Monte-Carlo tree search [69]). Combinatorial optimization has been researched extensively in healthcare settings [70], mainly to solve logistic tasks, resource allocation or help in global policy decisions, with a few examples in diagnosis and therapeutic decision tasks [71]. Among those methods, we will discuss exclusively the very active field of reinforcement learning, an approach worth mentioning both because of its recent super-human results on many non-medical tasks, as well as its unresolved limitations in highly regulated domains such as medicine.

Reinforcement Learning for Sequential Decision-Making

Reinforcement learning (RL) describes a family of methods inspired by operant conditioning, a form of associative learning through which rewards and punishments reinforce behaviors.

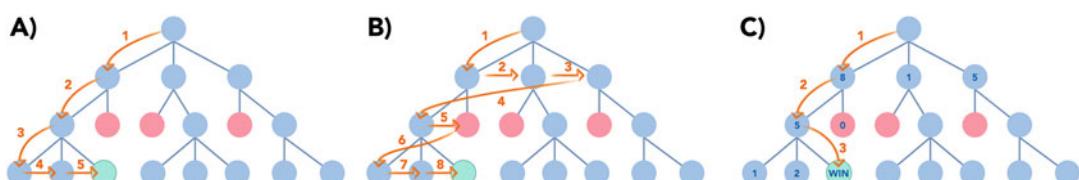


Fig. 7 Three search algorithms are looking for the target state (green). Depth-First (a) and Breadth-First Search (b) are brute force search algorithms, as compared to Best-

First Search (c), which is an informed search algorithm using *state values* to speed up search and avoid costly states (red)

They aim to solve sequential decision problems in settings where the underlying system and its dynamics are usually only partially known or too complex to simulate in reasonable time [72]. In order to facilitate the search for an optimal decision sequence, RL algorithms learn the value of states and/or actions through exploration, which allows at test-time to efficiently perform a best-first search approach (*model-based*) or to use a deterministic policy (*model-free*).

Recently, these approaches have greatly benefited from advances in Deep Learning, which unlocked the ability to represent and solve problems of much greater complexity. These *Deep* RL methods were particularly impressive in solving complex games, such as Go, Chess, Starcraft, Dota, several 1980s' Atari games, as well as many hard robotic tasks [73]. RL relies on a model called a Markov Decision Process, which formally describes a system's evolution in terms of state and actions, as well as a preference over certain states quantified. The Markov property specifies that the future only depends on the current state and the actions of the agent(s). This setting is pervasive in healthcare, for example, when planning a series of questions and tests to diagnose a patient or when defining clinical pathways for managing a particular disease.

Contrary to supervised learning approaches, in RL the goal is to learn a function able to map a state to an action (the *policy*) so that the average reward is maximized. Moreover, the data used in RL is usually gathered progressively through interactions rather than given a priori, and the history of actions leading to a particular state is used to learn how to behave optimally. Crucially, outside of particular settings, it would be dangerous, unethical and relatively slow to train agents in real environments. This is why researchers have relied either on simulated environments or on observational datasets (given this time a priori) where the actions are performed “off-policy” by real agents (e.g., doctors) rather than the algorithm itself.

In medicine, those methods have been used to train agents to perform active diagnosis [74] and triage [75] from structured evidence (see [76] for a formal treatment). It has also been used for

diagnosis using unstructured data, either directly or with a human-in-the-loop approach [77]. For example, it has been used to aid radiologists to parse medical images [78] and to perform diagnosis using new sources of data (e.g., using connected devices to diagnose lung cancer [79]). Regarding treatment, RL has been used in specific diseases and with a narrow set of actions to optimize. In particular, it has been used to learn the best treatment policy, or to adapt treatment, for diabetes [80], intensive care [81], sepsis [82] and respiratory failure [83].

Several drawbacks of the method impede the deployment of RL agents in real-life settings [84]. Mainly, the difficulty in measuring all potential confounders to avoid biased decisions, the high sensitivity to task parameters, and the opacity and the difficulty enforcing the safety of the learned policy. Methods have been proposed to help reduce the bias, produce human-understandable policies [85], evaluate the safety [86], and quantify the uncertainty of the resulting policies [87]. While safety could be improved using human-in-the-loop approaches, or logic-based constraints baked inside the environments (e.g., treatment dose limit), safe reinforcement learning remains a marginal but growing area of research. Like any ML-based system, deploying RL agents in real-life settings should be performed with extreme care only after a thorough clinical evaluation.

Bayesian Models for Decision Support

While fuzzy logic methods introduced in section “[Beyond First-Order Logic](#)” offer a path to working with partial and ambiguous knowledge, Bayesian methods take this a step further, offering a principled way of making decisions that take into account uncertainty and degrees of belief. Bayesian models can be learned from data, elicited from experts, or a combination of the two [88]. For example, the optimal model parameters can be identified by maximum a posteriori estimation,

$$\theta_{\text{MAP}} = \underset{\theta}{\operatorname{argmax}} P(D|\theta)P(\theta) \quad (3)$$

Carefully selecting the prior $P(\theta)$ allows domain knowledge to be injected into the learning process. While regression-based models described in section “[Learning from Data](#)” predict values of Y from features X , the outputs of Bayesian models are probability distributions $P(Y|X; \theta)$, supporting inferences on the statistical relationships between variables and principled uncertainty estimates for predictions. Despite these benefits, principled and effective priors can be hard to construct [89], and the computational cost of performing inference at train and test time is typically much higher than the equivalent overheads for function approximation methods.

Bayesian models are either generative models, representing the full joint distribution $P(X, Y; \theta)$, or are discriminative models, representing the conditional distribution $P(Y|X; \theta)$. Examples of generative Bayesian models include probabilistic graphical models, Gaussian mixture models, and linear discriminant analysis. Examples of discriminative Bayesian models include Gaussian processes, logistic regression, and Bayesian neural networks [88]. A generative model can generate any discriminative model over its variables by applying Bayes rule and marginalization. Thus, generative models can “drop” and “swap” inputs and outputs without requiring learning a new model or imputing missing values. This is particularly useful in clinical decision tasks such as diagnosis, where missing features are common and where variables such as those modelling the presence of a disease can be both input evidence and a target variable. This power comes at a cost – learning the joint distribution is often far more challenging than learning the conditional distribution for a small subset of target variables. Hence, discriminative models often achieve better performance by focusing modelling and computational resources on a simpler learning objective.

Bayesian Networks for CDSS

Bayesian networks (BNs) are intuitive graphical representations of joint probability distributions that offer several key advantages over logic and regression-based methods while presenting their

own unique challenges [90]. BNs are generative models comprising of nodes situated in a directed acyclic graph (DAG), representing variables and their causal relations, respectively (Fig. 8). It is this injection of causal knowledge that endows BNs with some unique properties, making them particularly well-suited as clinical decision models. For one, the joint probability distribution factorizes with respect to the DAG, greatly reducing the computational cost of inference and learning compared to unfactorized representations [88]. It can also be exploited to learn BNs from expert opinion, or directly from data using constraints-based and score-based learning algorithms, or a combination of the two [91]. This ability to learn from expert opinion is particularly important in tasks where carefully annotated data is scarce or non-existent, as is often the case in clinical decision tasks where labels are often generated by expert elicitation.

This addition of causal knowledge not only expedites model learning and inference – it also enables BNs to perform inferences that are impossible with other methods. The encoding of causal and conditional independencies as graphical structure allows complicated independence relations such as d-separation to be recast as simple graphical criteria, and enables causal and counterfactual “what-if” inferences under rigorous and testable assumptions [92]. These causal inferences are vital for many clinical decision tasks including individual risk prediction [93] and diagnosis [94], and can be used to construct explanations for automated decisions [95]. Finally, this graphical structure provides a route for extending BNs to influence diagrams by including decision and utility nodes [96]. These models are widely used in CDSS for determining optimal decisions via cost-utility analysis, including identifying optimal tests and evidence gathering through value of information (VOI) analysis [97].

For these reasons, BNs have been widely adopted as clinical decision support models, appearing across a range of medical specialisms including oncology [98], cardiology [99], neurology [100], psychiatry [101], and pediatrics [102], intensive care [103], and primary care [94], including systems developed by an emerging market of

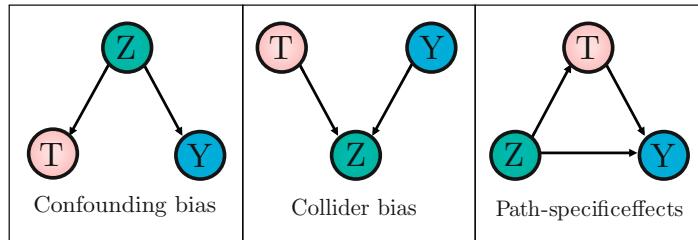


Fig. 8 Figure depicts common causal biases. Confounding bias arises when there is an unobserved common cause (Z) of both the treatment (T) and outcome (Y) variables. For example in [116] confounding bias is discussed as an explanation for the obesity Paradox, as smokers tend to have reduced rates of obesity and increased risk of morbidity. This bias can be corrected for using methods including the back-door adjustment formula and propensity score matching. Collider bias arises when the treatment (T) and outcome (Y) both independently cause a third variable (collider Z), which is then inappropriately controlled for either through statistical analysis or study design (selection bias). For example in [117], collider bias was identified as the cause of a strong association between locomotor disease and respiratory disease, spuriously suggesting that treating locomotor diseases could reduce the incidence of respiratory diseases. As the study was performed on hospitalized patients, and both locomotor and respiratory diseases independently cause hospitalization, then

selecting for hospitalized patients resulted in a spurious association between these disease groups, despite no such association being present in the general population. Collider bias can be controlled for using the tools of causal modelling and study design [118]. Path-specific effects arise in the presence of multiple causal pathways between treatment and outcome. For example in [115] several machine learning methods were employed to predict mortality of patients hospitalized for pneumonia. The models incorrectly learned to treat asthma (Z) as a protective risk-factor for pneumonia inpatients, reducing the mortality risk (Y). This result is due to path-specific effects – patients with asthma were more likely to receive aggressive treatments (T), which results in a lower mortality rate due to the path $Z \rightarrow T \rightarrow Y$, despite the negative direct causal effect of asthma on mortality from the path $Z \rightarrow Y$. Methods for resolving and correcting for these direct and indirect causal effects have been developed using counterfactual inference and mediation analysis [119]

digital healthcare companies [104, 105]. They have been applied to a variety of decision tasks, such as prognosis and risk prediction [106], treatment selection [107], and diagnosis [38, 94]. Despite these benefits, the use of BNs as decision support models has tapered off in recent years in favor of data-driven machine-learning algorithms. This transition has been driven by the move towards large, unstructured data sets and away from the expensive and time-consuming elicitation of structured expert knowledge. Another disadvantage of BNs in the big data regime is that the search space of DAGs grows super-exponentially with the number of variables. This makes learning BNs from large multivariate data sets very challenging, although there has been significant progress to reduce learning complexity in recent years [108]. On top of this, regression-based algorithms typically achieve higher accuracies than BNs at prediction tasks [109], and have emerged as the clear favorite models for constructing CDSS when sufficient labelled data is available. There is

however one area that Bayesian networks continue to play a key role – identifying cause and effect.

The Need for Causality in Clinical Decision-Making

The ultimate export of clinical decision support systems is to help clinicians select the best actions – to answer questions such as ‘what would happen to the patient if I administered this treatment?’. In statistical theory, interventions are treated distinctly from observational data as they carry additional information about cause and effect [92]. Consequently, the effects of interventions cannot be determined from observational data alone, requiring interventional data from randomized control trials or the application of causal inference techniques, which in turn require domain knowledge and causal assumptions that are not included in most clinical data sets [110]. Despite this, observational data sets such

as electronic health records are routinely used to make predictions and train decision algorithms without correcting for causal biases, potentially resulting in sub-optimal and even harmful decisions [93].

A classic example of this is the so-called obesity paradox – a routinely observed association between obesity and reduced mortality risk in patients with chronic diseases including cardiovascular disease, renal disease, diabetes, and stroke [111]. Causal explanations for this paradox have been proposed, including confounding bias [112] and collider bias [113], which in turn can be understood and mitigated using the tools of causal inference (Fig. 8). Decision support systems naively trained on observational data will learn to treat obesity as a protective risk factor for these conditions – a dangerous conclusion that could result in a less aggressive treatment regime being selected for obese patients despite their increased mortality risk. Similar examples arise in a variety of clinical settings [114, 115].

Causal biases are present to some extent in all real-world observational data sets [120]. The challenge is therefore to learn decision-making algorithms from abundant observational data while employing the tools of causal inference to correct for these biases. For example, the most common causal inference task in clinical decision-making is the estimation of treatment effects [121], which measure the response of individuals and populations to clinical interventions (treatments), and determine optimal personalized treatment decisions for individual patients [122]. Treatment effects quantify the causal relationship between a treatment variable T and an outcome Y (e.g., recovery rate) by comparing the expected value of the factual outcome (e.g., the recovery rate of treated individuals) to the corresponding counterfactual outcome (the recovery rate that would have been observed for treated individuals, if they had not been treated). This comparison of factual and counterfactual outcomes deals with the confounding bias that arises in the absence of randomization – patients with different propensities to receive treatment may have different recovery rates, e.g., sicker patients are more

likely to be treated and less likely to recover regardless of treatment (Fig. 9).

However, estimating treatment effects is notoriously difficult and typically relies on strong assumptions on how the data is generated and collected [124].

A common approach to estimating treatment effects is to apply the potential outcomes framework [125], under the assumption that all confounders between the treatment and outcome variable are observed. Recent advances extend this framework using methods from statistical machine learning including tree-based methods [126], deep learning [127], generative adversarial networks [128], and Gaussian processes [129]. Other advances have focused on applying the potential outcomes framework to more complicated prediction tasks involving time-series data, such as dynamic treatment regimens [130], dose-response curves [131], survival analysis [132], disease progression models [133], and state-of-the-art clinical decision support pipelines [49].

Structural causal models (SCMs) and Pearls' calculus of intervention are often used to estimate treatment effects in more complicated causal scenarios where fewer assumptions can be made [92, 134], such as in the presence of unobserved confounding or collider bias [135]. While these two approaches to causal inference – potential outcomes and SCMs – are complementary and logically equivalent [92] (pp. 2314), SCMs have the benefit of making causal assumptions explicit and testable by encoding them in directed acyclic graphs (DAGs), rather than relying on counterfactual assumptions such as strong ignorability [92, 134]. SCMs have been adopted in many cutting-edge methods, including estimating treatment effects using deep generative models [136] and dynamic treatment regimens [137]. Beyond estimating treatment effects, SCMs have been used to transport treatment effects between populations [138], to enhance explainability [139], fairness [140] in clinical decision-making, and in diagnostic decision-making [94], due to their ability to estimate the complicated counterfactuals required for diagnosis. These relatively recent advances hint that we are only beginning

Counterfactual	<p>Activities : Imagining, retrospection, explaining.</p> <p>Questions : How would my belief in Y change, given that I observe X to be some value, if instead I had forced X to take some different value?</p> <p>Examples : Is it likely that the patients symptoms were caused by this disease?</p>
Intervention	<p>Activities : Acting, intervening.</p> <p>Questions : How would my belief in Y change if I forced X to take some value?</p> <p>Examples : If we gave a patient a disease, how likely is it they will develop these symptoms?</p>
Association	<p>Activities : Observing, seeing.</p> <p>Questions : How would my belief in Y change given that I observe X to be some value?</p> <p>Examples : How often do patients with this disease develop these symptoms?</p>

Fig. 9 Pearl's causal hierarchy [123]

to understand the central role that causality will play in the next generation of clinical decision-making algorithms and, more broadly, artificial intelligence [141].

Explainability, Interpretability, and Fairness

One of the essential qualities of clinicians is their ability to communicate their findings and explain the reasoning behind their decisions. This ability allows for clinical decisions to be discussed and sense-checked against recommendations and clinical guidelines, and is critical for ensuring the trust and mutual understanding necessary for any automated system to work in a clinical setting. For

example, several deep learning systems for diagnosis from medical images have been found to learn image features relating to the specifics of the imaging device, which were predictive in a particular dataset but do not generalize and are clinically irrelevant [142]. Because these models are black-boxes, their decisions cannot be explained by humans, and these failures are often overlooked until they emerge in the field, when these systems fail to generalize to new clinical environments.

Explainable AI [143] seeks to deobfuscate these black-box models, typically by learning a second post-hoc model for generating explanations [139]. This approach is widely applied, especially for deep learning methods in clinical decision-making [144]. However, it is still debated whether or not these post-hoc

explanations are robust and reliable descriptions of what the black-box model is actually doing [142]. Another approach is to use models that are inherently interpretable and generate their own explanations [145]. However, many of these methods (including rules-based AI) ensure their interpretability through tight constraints that hamper model expressiveness, often resulting in decreased accuracy and leading some to propose a trade-off between interpretability and performance [146]. Human-in-the-loop training has also been proposed as a way to continuously sense-check and improve clinical decision-making systems by incorporating expert judgments in a continual training loop [147].

However, fully interpretable models can still produce undesirable decisions due to biases present in the training data. For example, in [148] it was found that a widely used commercial algorithm for guiding population health decisions exhibited racial bias. By using healthcare costs as a proxy to measure illness, the algorithm systematically underestimated the risk for African Americans. Historical biases in access to care inhibited the healthcare costs for African American patients, resulting in artificially lower risk predictions and lower priority access to care. Moreover, simply ignoring sensitive attributes is insufficient, as other attributes act as proxies (for example, postal code acting as a proxy for race) and can lead to even worse discrimination [149]. The emerging field of algorithmic fairness seeks to mitigate these biases [150], and while still in its infancy, it will be an essential part of AI research in the coming decade.

Conclusion

In this chapter, we have introduced the main approaches in artificial intelligence as applied to decision-making in healthcare. We have reviewed the recent innovations in AI, basic science, software and hardware products, as well as the maturing interface between technology and healthcare, and have argued that these developments indicate that the long-awaited paradigm shift in medical decision-making may be near. We have given a

functional classification of decision tasks that could be automated and successfully deployed at scale in the next decade. By introducing the main families of AI methods available, we aimed at highlighting their complementary properties and the potential of combined hybrid approaches. However, we have highlighted that these fast innovations, particularly in the machine learning field, do not come without risks and several biases that we have discussed. We presented causal methods to reduce causal biases in systems working from observational data and reiterated the need for modern clinical AI-product to prove their efficacy on clinical outcome metrics in Randomized Control Trials. Finally, we have argued that explainability, safety, and fairness are central to the applicability of AI to clinical decision-making beyond efficacy considerations and will undoubtedly continue to grow as a central research area in the next decade.

References

- Violato C. A brief history of the regulation of medical practice: Hammurabi to the national board of medical examiners. *J Sci Med.* 2016;2:122–122.
- Von Neumann J, Morgenstern O. Theory of games and economic behavior. Princeton University Press; 2007.
- Lehmann EL. Some principles of the theory of testing hypotheses. *Ann Math Stat.* 1950;21:1–26.
- Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis. *Science.* 1959;130(3366):9–21.
- De Dombal FT, Leaper DJ, Staniland JR, McCann AP, Horrocks JC. Computer-aided diagnosis of acute abdominal pain. *Br Med J.* 1972;2(5804):9–13.
- Spiegelhalter DJ, Knill-Jones RP. Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology. *J R Stat Soc: Ser A (General).* 1984;147(1):35–58.
- Shortliffe EH. Mycin: a knowledge-based computer program applied to infectious diseases. In: Proceedings of the annual symposium on computer application in medical care. American Medical Informatics Association; 1977. p. 66.
- Myers JD, Pople HE, Miller RA. Caduceus: a computerized diagnostic consultation system in internal medicine. In: Proceedings of the annual symposium on computer application in medical care. American Medical Informatics Association; 1982. p. 44.
- Ramnarayan P, Tomlinson AL, Kulkarni G, Rao A, Britto JF, et al. A novel diagnostic aid (ISABEL):

- development and preliminary evaluation of clinical performance. *Medinfo*. 2004;107:1091–5.
10. Barnett GO, Famiglietti KT, Kim RJ, Hoffer EP, Feldman MJ. Dexplain on the internet. In: Proceedings of the AMIA symposium. American Medical Informatics Association; 1998. p. 607.
 11. Imhoff M, Kuhls S. Alarm algorithms in critical care monitoring. *Anesth Analg*. 2006;102(5):1525–37.
 12. Bortolan G, Degani R, Willem JL. ECG classification with neural networks and cluster analysis. *Comput Cardiol*. 1991;20:177–80.
 13. Zoltan-Ford E, Chapanis A. What do professional persons think about computers? In: Use and impact of computers in clinical medicine. Springer; 1982. p. 51–67.
 14. Fieschi M, Dufour JC, Staccini P, Gouvernet J, Bouhaddou O. Medical decision support systems: old dilemmas and new paradigms? Tracks for successful integration and adoption. *Methods Inf Med*. 2003;42(3):190–8.
 15. Furman J, Seamans R. AI and the economy. *Innov Policy Econ*. 2019;19(1):161–91.
 16. Mali P, Yang L, Esveld KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. RNA-guided human genome engineering via cas9. *Science*. 2013;339(6121):823–6.
 17. Kirk D, et al. Nvidia cuda software and GPU parallel computing architecture. *ISMM*. 2007;7:103–4.
 18. Lupton D. Apps as artefacts: towards a critical perspective on mobile health and medical apps. *Societies*. 2014;4(4):606–22.
 19. Torous J, Lipschitz J, Ng M, Firth J. Dropout rates in clinical trials of smartphone apps for depressive symptoms: a systematic review and meta-analysis. *J Affect Disord*. 2020;263:413–9.
 20. Plante TB, O'Kelly AC, Macfarlane ZT, Urrea B, Appel LJ, Miller ER III, Blumenthal RS, Martin SS. Trends in user ratings and reviews of a popular yet inaccurate blood pressure-measuring smartphone app. *J Am Med Inform Assoc*. 2018;25(8):1074–9.
 21. Turakhia MP, Desai M, Hedlin H, Rajmane A, Talati N, Ferris T, Desai S, Nag D, Patel M, Kowey P, et al. Rationale and design of a large-scale, app-based study to identify cardiac arrhythmias using a smartwatch: the apple heart study. *Am Heart J*. 2019;207:66–75.
 22. Jain A, Way D, Gupta V, Gao Y, de Oliveira Marinho G, Hartford J, Sayres R, Kanada K, Eng C, Nagpal K, et al. Development and assessment of an artificial intelligence-based tool for skin condition diagnosis by primary care physicians and nurse practitioners in teledermatology practices. *JAMA Netw Open*. 2021;4(4):e217249.
 23. Almathami HKY, Win KT, Vlahu-Gjorgjevska E. Barriers and facilitators that influence telemedicine-based, real-time, online consultation at patients homes: systematic literature review. *J Med Internet Res*. 2020;22(2):e16407.
 24. Sadegh-Zadeh K, et al. Handbook of analytic philosophy of medicine. Springer; 2012.
 25. Saria S, Butte A, Sheikh A. Better medicine through machine learning: what's real, and what's artificial? *PLoS Med*. 2018;15:e1002721.
 26. Blum RL. Discovery, confirmation, and incorporation of causal relationships from a large time-oriented clinical data base: The RX project. *Comput Biomed Res*. 1982;15(2):164–87.
 27. Ekins S, Puhl AC, Zorn KM, Lane TR, Russo DP, Klein JJ, Hickey AJ, Clark AM. Exploiting machine learning for end-to-end drug discovery and development. *Nat Mater*. 2019;18(5):435.
 28. Haddad TC, Helgeson J, Pomerleau K, Makey M, Lombardo P, Coverdill S, Urman A, Rammage M, Goetz MP, LaRusso N. Impact of a cognitive computing clinical trial matching system in an ambulatory oncology practice. *J Clin Oncol*. 2018;36:6550.
 29. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AWR, Bridgland A, et al. Improved protein structure prediction using potentials from deep learning. *Nature*. 2020;577(7792):706–10.
 30. Friston KJ, Parr T, Zeidman P, Razi A, Flandin G, Daunizeau J, Hulme OJ, Billig AJ, Litvak V, Moran RJ, et al. Dynamic causal modelling of covid-19. *arXiv preprint arXiv:2004.04463*. 2020.
 31. Belthangady C, Royer LA. Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction. *Nat Methods*. 2019;16:1215–25.
 32. Smith JE, Russell RJ, Horne S. Critical decision-making and timelines in the emergency department. *BMJ Mil Health*. 2011;157(Suppl 3):S273–6.
 33. Tànfani E, Testi A. Advanced decision making methods applied to health care, vol. 173. Springer Science & Business Media; 2012.
 34. Fekom M, Vayatis N, Kalogeratos A. Dynamic epidemic control via sequential resource allocation. *arXiv preprint arXiv:2006.07199*. 2020.
 35. Felder S, Mayrhofer T, et al. Medical decision making. Springer; 2017.
 36. Erera S, Shmueli-Scheuer M, Feigenblat G, Nakash OP, Boni O, Roitman H, Cohen D, Weiner B, Mass Y, Rivlin O, et al. A summarization system for scientific documents. *arXiv preprint arXiv:1908.11152*. 2019.
 37. Fieschi M, Gouvernet J. Reasoning foundations of medical diagnosis revisited. *Yearb Med Inform*. 1999;8(01):78–82.
 38. Shwe MA, Middleton B, Heckerman DE, Henrion M, Horvitz EJ, Lehmann HP, Cooper GF. Probabilistic diagnosis using a reformulation of the internist-1/QMR knowledge base. *Methods Inf Med*. 1991;30(4):241–55.
 39. Zadeh LA. Information and control. *Fuzzy Sets*. 1965;8(3):338–53.
 40. Ahmadi H, Gholamzadeh M, Shahmoradi L, Nilashi M, Rashvand P. Diseases diagnosis using fuzzy logic methods: a systematic and meta-analysis review. *Comput Methods Prog Biomed*. 2018;161:145–72.
 41. Salmeron JL, Papageorgiou EI. A fuzzy grey cognitive maps-based decision support system for

- radiotherapy treatment planning. *Knowl-Based Syst.* 2012;30:151–60.
42. Papageorgiou EI. Fuzzy cognitive map software tool for treatment management of uncomplicated urinary tract infection. *Comput Methods Prog Biomed.* 2012;105(3):233–45.
 43. Sur RL, Dahm P. History of evidence-based medicine. *Indian J Urol: IJU: J Urol Soc India.* 2011;27(4):487.
 44. Oxman AD, Sackett DL, Guyatt GH, Browman G, Cook D, Gerstein H, Haynes B, Hayward R, Levine M, Nishikawa J, et al. Users' guides to the medical literature: I. How to get started. *JAMA.* 1993;270(17):2093–5.
 45. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al. Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 2020.
 46. Breiman L, et al. Statistical modeling: the two cultures. *Stat Sci.* 2001;16(3):199–231.
 47. Loh W-Y. Fifty years of classification and regression trees. *Int Stat Rev.* 2014;82(3):329–48.
 48. McCullagh P. Generalized linear models. Routledge; 2018.
 49. Jarrett D, Yoon J, Bica I, Qian Z, Ercole A, van der Schaar M. Clairvoyance: a pipeline toolkit for medical time series. In: International Conference on Learning Representations, 2020.
 50. Kline RB. Principles and practice of structural equation modeling. Guilford Publications; 2015.
 51. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. *IEEE Intell Syst Appl.* 1998;13(4):18–28.
 52. Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep learning, vol. 1. Cambridge: MIT Press; 2016.
 53. MacKay DJC. Information theory, inference and learning algorithms. Cambridge University Press; 2003.
 54. Lange S, Riedmiller M. Deep auto-encoder neural networks in reinforcement learning. In: The 2010 international joint conference on neural networks (IJCNN). IEEE; 2010. p. 1–8.
 55. Kingma DP, Welling M. Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114, 2013.
 56. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Advances in neural information processing systems; 2014. p. 2672–80.
 57. van den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K. Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499, 2016.
 58. Jimenez Rezende D, Mohamed S. Variational inference with normalizing flows. arXiv preprint arXiv:1505.05770, 2015.
 59. van den Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018.
 60. Ross Quinlan J. Induction of decision trees. *Mach Learn.* 1986;1(1):81–106.
 61. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on computational learning theory; 1992. p. 144–52.
 62. Halevy A, Norvig P, Pereira F. The unreasonable effectiveness of data. *IEEE Intell Syst.* 2009;24(2): 8–12.
 63. Bishop CM, et al. Neural networks for pattern recognition. Oxford University Press; 1995.
 64. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev.* 1958;65(6):386.
 65. Nair V, Hinton GE. Rectified linear units improve restricted Boltzmann machines. ICML. 2010;
 66. Kreinovich VY. Arbitrary nonlinearity is sufficient to represent all functions by neural networks: a theorem. *Neural Netw.* 1991;4(3):381–3.
 67. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G, Thrun S, Dean J. A guide to deep learning in healthcare. *Nat Med.* 2019;25(1):24–9.
 68. Knuth DE, Moore RW. An analysis of alpha-beta pruning. *Artif Intell.* 1975;6(4):293–326.
 69. Coulom R. Efficient selectivity and backup operators in Monte-Carlo tree search. In: International conference on computers and games. Springer; 2006. p. 72–83.
 70. Rais A, Viana A. Operations research in healthcare: a survey. *Int Trans Oper Res.* 2011;18(1):1–31.
 71. Iakovidis DK, Papageorgiou E. Intuitionistic fuzzy cognitive maps for medical decision making. *IEEE Trans Inf Technol Biomed.* 2010;15(1):100–7.
 72. Sutton R, Barto A. Reinforcement learning. MIT Press; 2018.
 73. Akkaya I, Andrychowicz M, Chociej M, Litwin M, McGrew B, Petron A, Paino A, Plappert M, Powell G, Ribas R, et al. Solving Rubik's cube with a robot hand. arXiv preprint arXiv:1910.07113, 2019.
 74. Kao HC, Tang KF, Chang EY. Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning. In: 32nd AAAI conference on artificial intelligence, AAAI 2018; 2018. p. 2305–13.
 75. Buchard A, Bouvier B, Prando G, Beard R, Livieratos M, Busbridge D, Thompson D, Richens J, Zhang Y, Baker A, et al. Learning medical triage from clinicians using deep q-learning. arXiv preprint arXiv:2003.12828, 2020.
 76. Araya M, Buffet O, Thomas V. Active diagnosis through information-look ahead planning. *Journées Francophones de Planification, Décision et Apprentissage pour la conduite de systèmes.* 2013.
 77. Lakdashti A, Ajorloo H. Content-based image retrieval based on relevance feedback and reinforcement learning for medical images. *ETRI J.* 2011;33(2):240–50.
 78. Sahba F, Tizhoosh HR, Salama MMA. A reinforcement learning framework for medical image segmentation. In: IEEE international conference on neural networks – conference proceedings. Institute of Electrical and Electronics Engineers; 2006. p. 511–7.

79. Liu Z, Yao C, Yu H, Taihua W. Deep reinforcement learning with its application for lung cancer detection in medical Internet of Things. *Futur Gener Comput Syst.* 2019;97:1–9.
80. Tejedor M, Woldaregay AZ, Godtliebsen F. Reinforcement learning application in diabetes blood glucose control: a systematic review. *Artif Intell Med.* 2020;104:101836.
81. Nemati S, Ghassemi MM, Clifford GD. Optimal medication dosing from suboptimal clinical examples: a deep reinforcement learning approach. In: 2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE; 2016. p. 2978–81.
82. Saria S. Individualized sepsis treatment using reinforcement learning. *Nat Med.* 2018;24(11):1641–2.
83. Prasad N, Cheng LF, Chivers C, Draugelis M, Engelhardt BE. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. In: Uncertainty in artificial intelligence – proceedings of the 33rd conference, UAI 2017. Corvallis: AUAI Press; 2017.
84. Lu MY, Shahn Z, Sow D, Doshi-Velez F, Lehman L-WH. Is deep reinforcement learning ready for practical applications in health-care? A sensitivity analysis of duel-DDQN for sepsis treatment. arXiv preprint arXiv:2005.04301, 2020.
85. Mansouri-poor F, Asadi S. Development of a reinforcement learning-based evolutionary fuzzy rule-based system for diabetes diagnosis. *Comput Biol Med.* 2017;91:337–52.
86. Gottesman O, Johansson F, Komorowski M, Faisal A, Sontag D, Doshi-Velez F, Celi LA. Guidelines for reinforcement learning in healthcare. *Nat Med.* 2019;25(1):16–8.
87. Aaron Sonabend W, Lu J, Celi LA, Cai T, Szolovits P. Expert-supervised reinforcement learning for offline policy learning and evaluation. arXiv e-prints, pages arXiv–2006, 2020.
88. Barber D. Bayesian reasoning and machine learning. Cambridge University Press; 2012.
89. Johnson SR, Tomlinson GA, Hawker GA, Granton JT, Feldman BM. Methods to elicit beliefs for Bayesian priors: a systematic review. *J Clin Epidemiol.* 2010;63(4):355–69.
90. Pearl J. Bayesian networks: A model of self-activated memory for evidential reasoning. In: Proceedings of the 7th conference of the cognitive science society, University of California, Irvine, CA, USA; 1985. p. 15–7.
91. Koller D, Friedman N. Probabilistic graphical models: principles and techniques. MIT Press; 2009.
92. Pearl J. Causality. Cambridge University Press; 2009.
93. Prosperi M, Guo Y, Sperrin M, Koopman JS, Min JS, He X, Rich S, Wang M, Buchan IE, Bian J. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nat Mach Intell.* 2020;2(7):369–75.
94. Richens JG, Lee CM, Johri S. Improving the accuracy of medical diagnosis with causal machine learning. *Nat Commun.* 2020;11(1):1–9.
95. Miller T. Explanation in artificial intelligence: insights from the social sciences. *Artif Intell.* 2019;267:1–38.
96. Kjaerulff UB, Madsen AL. Bayesian networks and influence diagrams, vol. 200. Springer Science+ Business Media; 2008. p. 114.
97. Nielsen TD, Jensen FV. Bayesian networks and decision graphs. Springer Science & Business Media; 2009.
98. Senen MB, Nicholson AE, Banares-Alcantara R, Kadir T, Brady M. Bayesian networks for clinical decision support in lung cancer care. *PLoS One.* 2013;8(12):e82349.
99. de Oliveira LSC, Andreão RV, Sarcinelli-Filho M. Premature ventricular beat classification using a dynamic Bayesian network. In: 2011 annual international conference of the IEEE engineering in medicine and biology society. IEEE; 2011. p. 4984–7.
100. Ding X, Bucholc M, Wang H, Glass DH, Wang H, Clarke DH, Bjourson AJ, Le Roy CD, O’Kane M, Prasad G, et al. A hybrid computational approach for efficient Alzheimers disease classification based on heterogeneous data. *Sci Rep.* 2018;8(1):1–10.
101. Curiac D-I, Vasile G, Banias O, Volosencu C, Albu A. Bayesian network model for diagnosis of psychiatric diseases. In: Proceedings of the ITI 2009 31st international conference on information technology interfaces. IEEE; 2009. p. 61–6.
102. Dexheimer JW, Abramo TJ, Arnold DH, Johnson K, Shyr Y, Ye F, Fan K-H, Patel N, Aronsky D. Implementation and evaluation of an integrated computerized asthma management system in a pediatric emergency department: a randomized clinical trial. *Int J Med Inform.* 2014;83(11):805–13.
103. Schurink CAM, Lucas PJF, Hoepelman IM, Bonten MJM. Computer-assisted decision support for the diagnosis and treatment of infectious diseases in intensive care units. *Lancet Infect Dis.* 2005;5(5):305–12.
104. Razzaqi S, Baker A, Perov Y, Middleton K, Baxter J, Mullarkey D, Sangar D, Taliercio M, Butt M, Majeed A, et al. A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis. arXiv preprint arXiv:1806.10698, 2018.
105. Lin N, Lu C, Liu N, Liu J. Mandy: Towards a smart primary care chatbot application. In: International symposium on knowledge and systems sciences. Springer; 2017. p. 38–52.
106. Stojadinovic A, Bilchik A, Smith D, Eberhardt JS, Ward EB, Nissan A, Johnson EK, Protic M, Peoples GE, Avital I, et al. Clinical decision support and individualized prediction of survival in colon cancer: Bayesian belief network model. *Ann Surg Oncol.* 2013;20(1):161–74.

107. Jiang X, Wells A, Brufsky A, Neapolitan R. A clinical decision support system learned from data to personalize treatment recommendations towards preventing breast cancer metastasis. *PLoS One.* 2019;14(3):e0213292.
108. Zheng X, Aragam B, Ravikumar PK, Xing EP. DAGs with NO TEARS: continuous optimization for structure learning. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18). Curran Associates Inc., Red Hook, NY, USA, 2018;9492–9503.
109. Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd international conference on Machine learning (ICML '06). Association for Computing Machinery, New York, NY, USA, 2006;161–168. <https://doi.org/10.1145/1143844.1143865>.
110. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ.* 2018;361.
111. Romero-Corral A, Montori VM, Somers VK, Korinek J, Thomas RJ, Allison TG, Mookadam F, Lopez-Jimenez F. Association of bodyweight with total mortality and with cardiovascular events in coronary artery disease: a systematic review of cohort studies. *Lancet.* 2006;368(9536):666–78.
112. Preston SH, Stokes A. Obesity paradox: conditioning on disease enhances biases in estimating the mortality risks of obesity. *Epidemiology (Cambridge, Mass).* 2014;25(3):454.
113. Banack HR, Kaufman JS. The obesity paradox explained. *Epidemiology.* 2013;24(3):461–2.
114. Lucero RJ, Lindberg DS, Fehlberg EA, Bjarnadottir RI, Li Y, Cimiotti JP, Crane M, Prosperi M. A data-driven and practice-based approach to identify risk factors associated with hospital-acquired falls: applying manual and semi-and fully-automated methods. *Int J Med Inform.* 2019;122:63–9.
115. Cooper GF, Aliferis CF, Ambrosino R, Aronis J, Buchanan BG, Caruana R, Fine MJ, Glymour C, Gordon G, Hanusa BH, et al. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artif Intell Med.* 1997;9(2):107–38.
116. Stokes A, Preston SH. Smoking and reverse causation create an obesity paradox in cardiovascular disease. *Obesity.* 2015;23(12):2485–90.
117. Sackett DL. Bias in analytic research. In: The case-control study consensus and controversy. Elsevier; 1979. p. 51–63.
118. Bareinboim E, Pearl J. Controlling selection bias in causal inference. In: Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, in Proceedings of Machine Learning Research 2012;22:100–108. Available from <https://proceedings.mlr.press/v22/bareinboim12.html>.
119. Pearl J. Direct and indirect effects. arXiv preprint arXiv:1301.2300, 2013.
120. Grimes DA, Schulz KF. Bias and causal associations in observational research. *Lancet.* 2002;359(9302):248–52.
121. Shpitser I, Pearl J. Identification of conditional interventional distributions. arXiv preprint arXiv:1206.6876, 2012.
122. Bica I, Alaa AM, Lambert C, van der Schaar M. From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. *Clin Pharmacol Ther.* 2020;109:87.
123. Pearl J. The seven tools of causal inference, with reflections on machine learning. *Commun ACM.* 2019;62(3):54–60.
124. Guo R, Cheng L, Li J, Hahn PR, Liu H. A survey of learning causality with data: problems and methods. *ACM Comput Surv (CSUR).* 2020;53(4):1–37.
125. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol.* 1974;66(5):688.
126. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. *J Am Stat Assoc.* 2018;113(523):1228–42.
127. Johansson F, Shalit U, Sontag D. Learning representations for counterfactual inference. In: International conference on machine learning. PMLR, 2016.
128. Yoon J, Jordon J, van der Schaar M. Ganite: estimation of individualized treatment effects using generative adversarial nets. In: International conference on learning representations; 2018.
129. Alaa AM, van der Schaar M. Bayesian inference of individualized treatment effects using multi-task Gaussian processes. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 2017;3427–3435.
130. Murphy SA. Optimal dynamic treatment regimes. *J R Stat Soc: Ser B (Stat Methodol).* 2003;65(2):331–55.
131. Schwab P, Linhardt L, Bauer S, Buhmann JM, Karlen W. Learning counterfactual representations for estimating individual dose-response curves. In: Proceedings of the AAAI Conference on Artificial Intelligence 2020;34(04):5612–5619. <https://doi.org/10.1609/aaai.v34i04.6014>.
132. Chapfuwa P, Assaad S, Zeng S, Pencina M, Carin L, Henao R. Survival analysis meets counterfactual inference. arXiv preprint arXiv:2006.07756, 2020.
133. Schulam P, Saria S. Reliable decision support using counterfactual models. In: Advances in neural information processing systems; Curran Associates, Inc. Red Hook, NY, USA, 2017. p. 1697–708.
134. Pearl J. The foundations of causal inference. *Sociol Methodol.* 2010;40(1):75–149.
135. Shpitser I, VanderWeele T, Robins JM. On the validity of covariate adjustment for estimating causal effects. In: Proceedings of the twenty-sixth conference on uncertainty in artificial intelligence, UAI'10, page 527536, Arlington, Virginia, USA. AUAI Press; 2010.

136. Louizos C, Shalit U, Mooij JM, Sontag D, Zemel R, Welling M. Causal effect inference with deep latent-variable models. In: Advances in neural information processing systems; La Jolla, CA: Neural Information Processing Systems, 2017. p. 6446–56.
137. Zhang J, Bareinboim E. Near-optimal reinforcement learning in dynamic treatment regimes. In: Advances in neural information processing systems; Curran Associates, Inc, 2019. p. 13401–11.
138. Pearl J, Bareinboim E. External validity: from do-calculus to transportability across populations. *Stat Sci.* 2014;29:579–95.
139. Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harv J Law Technol.* 2017;31:841.
140. Pfohl S, Duan T, Ding DY, Shah NH. Counterfactual reasoning for fair clinical risk prediction. arXiv preprint arXiv:1907.06260, 2019.
141. Pearl J. Theoretical impediments to machine learning with seven sparks from the causal revolution. arXiv preprint arXiv:1801.0f016, 2018.
142. Lapuschkin S, Waldchen S, Binder A, Montavon G, Samek W, Müller K-R. Unmasking clever Hans predictors and assessing what machines really learn. *Nat Commun.* 2019;10(1):1–8.
143. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Garcia S, Gil-López S, Molina D, Benjamins R, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion.* 2020;58:82–115.
144. Holzinger A, Langs G, Denk H, Zatloukal K, Muller H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev: Data Min Knowl Disc.* 2019;9(4):e1312.
145. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell.* 2019;1(5):206–15.
146. Gunning D. Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency (DARPA), ND Web. 2017. 2(2).
147. Holzinger A. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Inform.* 2016;3(2):119–31.
148. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 2019;366 (6464):447–53.
149. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. ACM Digital Library; 2012. p. 214–26.
150. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635, 2019.



Artificial Intelligence for Medical Diagnosis

12

Jonathan G. Richens and Albert Buchard

Contents

Introduction	182
Diagnostic Reasoning	182
Knowledge-Based Diagnosis	185
Rule-Based Diagnosis	186
Fuzzy-Logic Systems	186
Ontology-Based Systems	187
Model-Based Diagnosis	188
Abductive Diagnosis	188
Bayesian Diagnosis	189
Causal Reasoning for Diagnosis	191
Machine Learning for Diagnosis	191
Learning from Data	191
Machine Learning as Function Approximation	193
Three Supervised Methods	194
Other Machine Learning Formalisms for Diagnosis	195
The Importance of Data	196
Outlook	196
References	197

Abstract

Medical diagnosis has been one of the primary targets of Artificial Intelligence research since the inception of the field. In recent years, rapid

advances in Artificial Intelligence have seen the emergence of diagnostic algorithms that perform as well as clinicians and can be applied at scale in clinical practice. This chapter presents a broad picture of the foundations, history, and the current state of AI in medical diagnosis. We provide an overview of the complex and interdependent tasks required to perform diagnosis and explore how ideas from the study of diagnostic reasoning and diagnostic errors can guide the effective development and deployment of Artificial Intelligence solutions.

J. G. Richens (✉)
AI Research, Babylon Health, London, UK
e-mail: jonathan.richens@babylonhealth.com

A. Buchard
Service de Psychiatrie adulte, Hopitaux Universitaires de Genve, Geneva, CH, Switzerland

We then review the three main approaches to diagnostic AI; rules-based, model-based, and machine learning, detailing their strengths and weaknesses, and how each of these approaches tackles diagnosis from a different angle.

Keywords

Clinical decision-making · Artificial intelligence · Medical diagnosis · Causality · Computer-assisted diagnosis

Introduction

Diagnosis is one of the most important tasks that clinicians perform. Diagnostic errors account for one of the largest sources of burden to healthcare systems globally [1], with wrong or delayed diagnoses occurring in an estimated 10–15% of cases [2]. These errors cause more serious patient harm than any other type of medical error [3], resulting in between 40,000 and 80,000 death per year in US hospitals alone [1]. Improving the accuracy and timeliness of diagnoses reduces costs incurred by inappropriate testing, wrong treatments, and malpractice lawsuits, with the potential to save up to \$100 billion per year in the USA alone [4]. Improving access to clinicians for undeserved communities is also an important component of any solution to these issues, and is increasingly prioritized by governments, NGO's, and digital health companies [5]. Over the past 40 years, AI has always seemed only a few years away from resolving these issues; promising to improve access through automation and digitization, while simultaneously reducing the incidence of diagnostic errors.

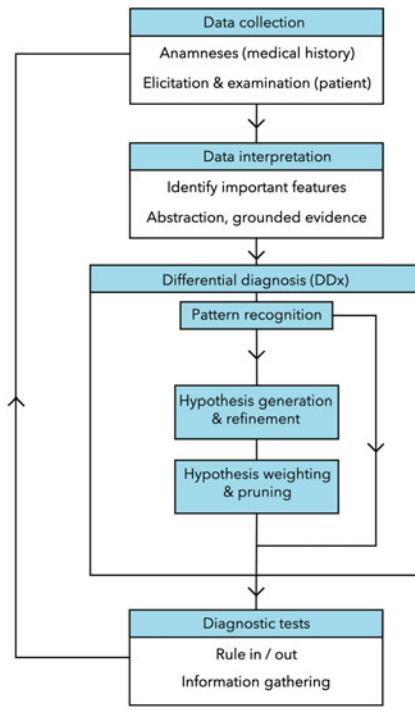
In this chapter we present a broad picture of the foundations, history and current state of AI for diagnosis. Diagnosis is a specialized instance of clinical decision making, and in this chapter we extend upon the methods presented in ► Chap. 11, “[Artificial Intelligence for Medical Decisions](#),” with a focus on the specific complexities that arise in diagnostic tasks. We provide a brief review of diagnostic reasoning, including the non-systemic biases and limitations that cause

diagnostic errors, how these can be addressed with AI, and how they can inform the development and deployment of AI solutions. We will cover the three main approaches to diagnostic AI: rule-based, model-based, and machine learning. These three approaches developed at different times and were borne out of fundamentally different AI paradigms, and each tackles diagnosis from a different angle. Due to the complexity of the diagnostic process, which comprises of multiple sub-tasks with their own requirements and strategies, none of these approaches alone is sufficient to characterize and “solve” diagnosis, which ultimately requires a hybrid of these methodologies.

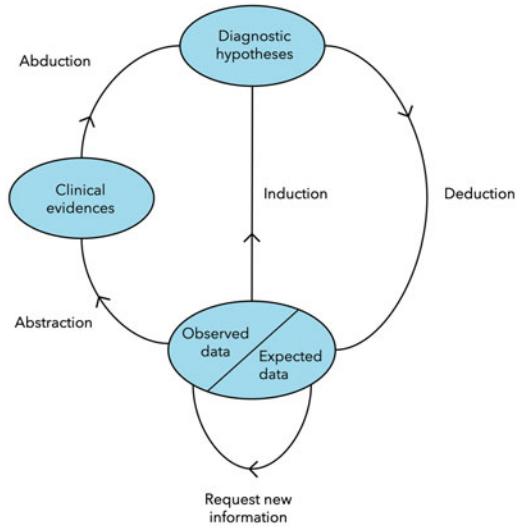
Diagnostic Reasoning

The goal of diagnosis is to identify the possible diseases or conditions that best explain the signs and symptoms presented by the patient. In clinical practice, diagnosis is an ongoing dynamic process between the patient and doctor, consisting of a sequence of interdependent tasks that revolve around gathering and interpreting patient data, and generating and testing hypotheses (Fig. 1). In their initial interaction, the doctor collects relevant medical history (anamnesis) and symptoms through direct elicitation from the patient, gathers the patient’s signs through physical examination, and optionally plans para-clinical tests (e.g., lab tests or imagery) [6]. This evidence is then interpreted and used to generate and weight diagnostic hypotheses, which are then used to inform the gathering of additional information to rule in or out conditions. Once these hypotheses are whittled down to a final diagnosis, remedial actions are selected, for example selecting treatments or “wait and see.”

We refer to this dynamic and investigative process, through which the underlying cause of the patient’s presentation is discovered, as “active diagnosis” [6, 7]. The iterative process of hypothesis generation and hypothesis testing has been formally described by Pierce’s “Select and Test model” [8], later adapted to medical diagnosis by Ramoni (Fig. 1) [9, 10]. This describes diagnostic reasoning in terms of sub-tasks involving



[a]



[b]

Fig. 1 (a) Depiction of active diagnosis process. Data is gathered, interpreted and abstracted. If the diagnostic case is identified in memory (pattern recognition), diagnostic tests can be immediately selected (type 1 reasoning), otherwise a process of hypothesis generation and refinement is initiated (type 2 reasoning). Diagnostic tests feed back into data collection, and the process continues until a final diagnosis is reached. (b) The Select and Test model [8, 9] describes the diagnosis process as first an abstraction step,

abstraction, abduction, deduction and induction which, when performed by clinicians, require a combination of complex cognition, metacognition, experience and formal knowledge [11]. These cognitive processes and their relation to each other are often described using the dual-process theory of clinical reasoning [12], which frames diagnosis as a combination of automatic, intuitive reasoning based on past experiences (type 1 reasoning), and conscious, reflective and analytic reasoning (type 2 reasoning) based on logic and formal knowledge [13]. Type 1 reasoning is responsible for generating diagnostic hypotheses, and is based on similarity matching the current diagnostic case with past experiences and exemplars. Type 2 reasoning is employed if a

converting raw observations into useful evidence, followed by three types of logical inference steps: abduction as the process of hypothesis generation from evidence, the deduction of which new evidence to expect given the hypotheses, which is used to inform information gathering, and potentially an induction step which allows to immediately revise hypotheses following unexpected observations, or to generate new knowledge

diagnosis cannot be identified from memory, or to correct for errors in type 1 reasoning. Type 2 reasoning also controls the search for information or the selection of follow-up tests which are not immediately triggered by a specific situation, as well as the weighting of diagnostic hypotheses [14]. It has even been suggested to employ symbolic logic [15].

Active diagnosis can be described as a function f that takes as input a set of medical evidence \mathcal{X} and returns a ranked set of disease hypotheses \mathcal{H} and a ranked set of diagnostic tests \mathcal{T} , which includes any follow-up and information gathering actions. On top of this, it should include a stopping criterion \mathcal{A} that determines whether to continue gathering information, or if there is

sufficient certainty to return a diagnosis. Additionally, clinicians are able to identify which cases can be solved with type 1 or type 2 reasoning, and the stopping criterion could be extended to include switching between different diagnostic algorithms. Formally,

$$f : \mathcal{X} \rightarrow \mathcal{O}^{|\mathcal{H}|}, \mathcal{O}^{|\mathcal{T}|}, \mathcal{A} \quad (1)$$

where \mathcal{O} any ordered set, such as \mathbb{R} or $[0,1]$. However, many diagnostic algorithms typically avoid replicating the difficult data-gathering process in its entirety. They instead focus on the task of “passive diagnosis” which simplifies diagnosis to a classification task and allows for simple learning objectives and performance measures to be defined. Consequently, passive diagnosis can be defined as a single-step process producing a ranked set of diagnostic hypotheses given a set of medical evidence:

$$f : \mathcal{X} \rightarrow \mathcal{O}^{|\mathcal{H}|} \quad (2)$$

For example, pattern recognition algorithms that are used to predict diagnostic labels from electronic health records [16] or medical images [17] are passive diagnosis algorithms, whereas the QMR/INTERNIST-1 system is an active diagnosis system as it includes information gathering heuristics and a stopping criterion [18].

The majority of diagnostic algorithms that employ modern machine learning techniques can be categorized as passive diagnosis algorithms [19]. While these approaches typically achieve the highest accuracies at diagnostic classification tasks, they often lack interpretability [20], and can struggle with error correction [21]. Learning algorithms are vulnerable to taking “shortcuts,” achieving high accuracies by exploiting confounding variables that subsequently prevent them from generalizing to real-world settings [22], as evidenced by deep-learning algorithms for diagnosing hip fractures [23] and Covid-19 [24] that have been found to make diagnoses based on confounding variables such as the timing or specific imaging machine used. Evidently, these algorithms are not intended to perform the full suite of diagnostic reasoning, but rather to

achieve high accuracy in rapid pattern recognition without introspection – i.e., type 1 reasoning. Model-based algorithms are better suited to solving diagnostic tasks involving explanation, information gathering, hypothesis testing, and causal reasoning – tasks which are associated with type 2 reasoning in clinicians. Finally, rules-based systems are ideally suited to carrying out the many clinical tasks that require fixed, interpretable, and deductive rules to be followed, such as those encoded in clinical guidelines. A key challenge is to bring these different approaches together to build diagnostic systems that can support all parts of active diagnosis, rather than focusing on a single sub-task, while ensuring that errors are identified and corrected and that clinical guidelines are followed.

Diagnostic algorithms are typically implemented as decision support systems, which aim to augment human decision making and prevent human errors [25]. While many comprehensive diagnostic support systems have been developed over the last 40 years, only the simplest systems catering to general alerts, reminders, summary dashboards and information retrieval systems have ever been deployed at scale [26]. This is due in large part to the more advanced systems often impeding on rather than improving clinical practice [27]. In order to develop diagnostic systems which have an impact in the real world, it is important to first understand the different types of diagnostic error and how they arise in clinical practice [28]. Firstly, knowing the most common causes of serious diagnostic errors helps to focus research efforts on the most impactful interventions. One example of this is satisfaction of search (SoS) bias, where a clinician prematurely terminates information gathering after identifying the first abnormality, and in doing so misses evidence that would otherwise change the diagnosis. Radiographic images are complex and difficult to visually search, and as a result SoS is one of the primary causes of error in diagnostic radiology, contributing to approximately 22% of misdiagnosed cases [29]. This motivates the deployment of computer vision algorithms for image segmentation, which can identify abnormalities with high precision in “blind spot” areas

of images [30]. Another major type of error in radiology is the failure to identify the underlying cause of an abnormality detected in the image, typically due to cognitive biases including anchoring bias, premature closure, and attribution bias [31] (contributing to 10% of errors). Image classifiers can accurately identify and differentiate abnormalities in radiographs, even surpassing experts [32]. The second motivation for understanding cognitive errors is that errors with different causes require different interventions, with every diagnostic error having on average 6 different root causes [33], each potentially requiring a different kind of intervention. For example, if a diagnostic classification task has a high error rate, but these errors are due to irreducible uncertainty (e.g., clinicians are approximately Bayes-optimal classifiers), algorithmic classifiers cannot be expected to have a significant impact, and efforts

should focus instead on improving data gathering and interpretation. Given the complexity of the active diagnosis task, and the number of interacting sources of diagnostic error that must be addressed, it is perhaps unsurprising that clinical decision support systems have struggled to achieve their intended impact. Figure 2 provides a non-exhaustive taxonomy of the main cognitive causes of diagnostic errors, examples of their prevalence in different diagnostic areas, and examples of current AI solutions.

Knowledge-Based Diagnosis

Because medical knowledge is vast, symbolic, and deeply structured, the most natural approach for automating medical diagnosis is through knowledge-based methods. Clinical practice is

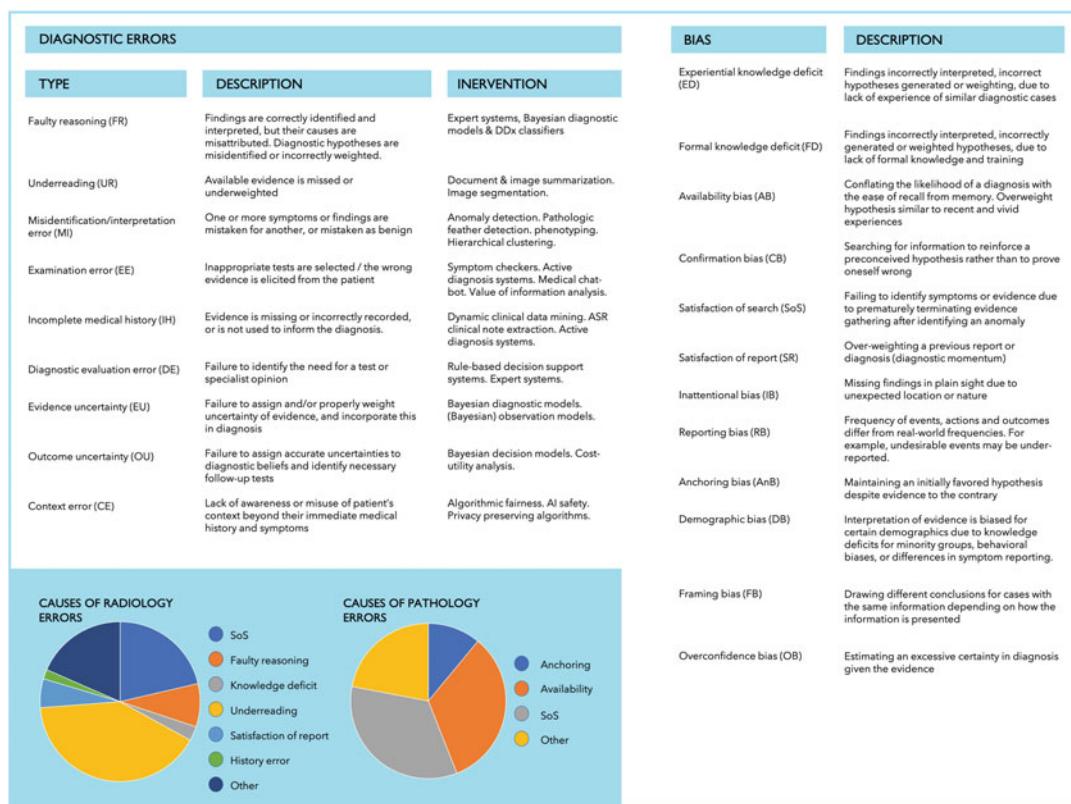


Fig. 2 Table shows nonexhaustive list of common errors in diagnosis, and current AI interventions. Table also describes common cognitive causes of these errors. Pie

charts depict the prevalence of different error types for diagnosis in radiology [29] and pathology [34]

highly codified through diagnosis criteria, staging scores, checklists, treatment decision algorithms, up to the detailed list of information a patient should legally receive. Whether it be about anatomy, physiology, pathological processes, or clinical practice, this type of knowledge is hard to learn from data, but readily represented in a logic-based format gathered from experts.

Knowledge-based methods exploit such logical statements stored in a database referred to as a “Knowledge-Base” (KB). While recent research has shown some of those KB can be learned from data [35], the process is not fully automated, and experts are often required to translate their knowledge into logical statements that the system can use. As described in the previous chapter, logic-based systems can be *crisp* or *fuzzy*, and depending on the type of knowledge stored in the KB, are referred to as rule-based or model-based systems.

Rule-Based Diagnosis

We call “rule-based” a system which relies on a collection of nested rules of the form “IF A THEN B” organized in complex decision trees. These systems are fully explicit and understandable by humans, and their behavior can be easily defined and debugged. They are particularly well suited for describing diagnosis rule such as:

IF

1. Rapid onset
2. AND Middle ear effusion
3. AND Otalgia interfering with sleep
4. AND Core temperature above 39 °C

THEN (Diagnose) Acute Severe Otitis Media.

The rule-based format is also the best suited to explicitly encode the natural flow of question a doctor may ask during Active Diagnosis, during which they aim to gather key information to improve his diagnosis. For example, the previous “passive” diagnostic rule could be associated with data gathering rule, in order to gather missing evidence:

IF Otalgia interfering with sleep *THEN*

1. ASK when the symptoms started
2. LOOK FOR Middle ear effusion
3. MEASURE Core temperature above 39 °C

ASK, LOOK FOR, and MEASURE, are three equivalent terms which mean the system should test for the presence of those signs, whether through human input or in an automated fashion. However, such approaches fail for large problems due to the sheer number of rules necessary to cover the space of valid states, which in part explains why the most prominent rule-based expert-systems developed in the 70–80’s each took on the order of a decade to complete. For example, the MYCIN system [36], the archetypal rule-based system, took six years to complete. It was deployed during that first wave of medical expert systems, along with many other rule-based systems such as PUFF [37], AI/RHEUM [38], or CADIAG-1 [39]. As discussed in the previous chapter, those systems mainly served as research prototypes and never were deployed at scale, in large part due to their incompatibility with a clinician’s daily practice.

Fuzzy-Logic Systems

Another dimension to rule-based systems is whether they are *crisp* or *fuzzy*. Crisp systems correspond to the classic logic formalism where statements are either True or False. *Fuzzy-logic* on the other hand relaxes this constraint allowing for statements to receive a *truth value* between 0 and 1. As described in the previous chapter, this approach is rooted in the notion of partial membership and fuzzy sets [40], and can deal with ambiguous situations and facilitates knowledge elicitation from experts in areas where medical knowledge is imperfect.

For diagnosis, these truth values can represent *uncertainty* and encode such statement as “Flu Rarely Causes Vomiting,” where “Vomiting” is a partial member to the set of symptoms caused by flu. They are also useful to encode *vague terms* which may be more natural for experts or represent the level of description matching with clinical

knowledge. For example, the term “early adulthood” is vague but clinically useful to talk about age groups in such statements as “Schizophrenia’s usually starts during early adulthood.” Each integer age value could then be associated with a partial membership to the “young adulthood” set, with a peak value around 25 years old. Both the rule-based and the ontological knowledge format can be fuzzified to allow for vagueness during knowledge elicitation and inference.

These methods have successfully been used since the 1980’s to perform diagnosis, in such systems as SPHYNX [41], CADIAGS-2 [39] or MILORD [42], and fuzzy-logic systems are still being actively researched [43]. Future systems may always require a layer of logic-based reasoning to ensure safety, and allow for the direct use of symbolic medical knowledge. Hybrid methods, such as Neuro-fuzzy systems combining fuzzy-logic with Neural Networks [44], offer the possibility to combine both approaches’ strength and build more efficient products.

Ontology-Based Systems

The second knowledge format is called ontological and encodes a model of the world. It describes symbolic entities, or concepts, and their associations with each other through relationships such as

*“Meningitis IS AN Infectious Disease” or
“Meningitis IS LOCATED IN THE Meninges”
or
“Meningitis CAUSES Fever”*

This second approach can store much of the medical knowledge, which is highly symbolic. It is also more manageable for large problems such as diagnosis, as the representation does not encode a specific step-by-step policy but relies on “model-based” inference algorithms to make decisions. For example, the KB could store that both Meningitis and Otitis Media causes Fever. However, the diagnostic algorithm would then have to rely on a specific strategy to either rank the two hypotheses (i.e., passive diagnosis) or ask for information to disambiguate them (i.e., active diagnosis). As detailed in the next section, these

diagnostic algorithms are called model-based algorithms because they perform abduction by exploiting a causal model of the world [45, 46].

Contrary to rule-based systems which are monolithic products, ontologies usually represent a general model of the world which can be shared and used in different applications. They can be exploited by logic-based [47], as well as learning-based algorithms to solve various tasks (e.g., link prediction [48], question answering [49], or ontology alignment [50]). And as detailed next, although they are not called an ontology, probabilistic methods also use causal models linking diseases to medical evidence for diagnosis.

Prominent ontologies such as ICD10 [51], SNOMED-CT [52], or UMLS [53] are extensively used in EHR systems to code for medical concepts. They allow for semantic interoperability, normalization of data gathering practice, and they facilitate research efforts [54]. However, those widely available ontologies focus on building a “class-system” in order to place concept in semantic hierarchies, such as

*“Meningitis IS AN Infectious Disease”
“Infectious Disease IS A Disease”*

This type of semantic knowledge has little use for an automated diagnosis system and, generally, those openly available ontologies do not hold any knowledge, or incomplete knowledge, about the relationships between diseases and symptoms.

On the other end of the spectrum, phenotype databases were developed to describe and link patient presentations, or phenotypes, to either a genotype or a particular disease. There have been several efforts in the 1970s–1980s to build such causal networks. For example, CASNET [55], MUNIN [56], and INTERNIST-1/CADUCEUS and its derivatives (e.g., QMR-DT) [45, 57] had such disease to symptom connections, but those systems are proprietary and now four decades old. More recent efforts such as Orphanet [58], or the Human Phenotype Ontology [59] are research-oriented, have focused on rare diseases, and the data is only available under certain conditions. The most comprehensive KB might be Google’s Health Knowledge Graph [60], although little is

known about its coverage and it most certainly will remain proprietary.

Overall, the field has yet to see an entirely open ontology which can support standard diagnosis tasks. Efforts such as WikiData for medicine [61] or Human Dx [62] made promises of an open and rich ontology supporting clinical tasks, but they are still incomplete or unavailable to the public. The lack of such a system is a testament to the complexity of the task. However, in the future, open ontologies of common clinical situations and national and international guidelines will be essential to build public benchmarks and safe systems that can inter-operate with clinicians. In the next section, we describe model-based algorithms which operate on causal knowledge about diseases to perform diagnosis and present their advantages over rule-based systems.

Model-Based Diagnosis

Supplanting early probabilistic approaches to diagnosis [63], the first large scale diagnostic systems developed were logic-based expert systems [36, 45]. As described in the previous section, these systems classified patients following a pre-defined diagnostic path from their symptoms and signs to a diagnostic decision, based on the application of deductive rules [64]. Despite initial optimism that all medical decision tasks could be solved in this way, cracks started to emerge in this approach in the 1980s [65]. For one, logical proofs have a limited applicability to solving problems with dynamic, incomplete, ambiguous or unstructured data, making these systems inherently brittle to many real-world diagnostic tasks. On top of this, the labor required to elicit expert knowledge and maintain and update large rule-based knowledge bases became a major bottleneck in the construction and deployment of expert systems [66].

Model-based methods started to be investigated as an alternative approach to expert systems in the late 1970s [67]. The idea behind these initial efforts was to exploit objective information about a system's behavior, rather than relying on expert knowledge based on subjective experience to

model the diagnostic process itself. The major insight of this approach was to separate domain knowledge from diagnostic rules, and instead to allow diagnostic rules to emerge from a combination of a diagnostic model and a set of simple inference rules and hypothesis ranking heuristics [68]. Even relatively small models including a handful of variables can in principle encode a vast number of rules which are defined for every possible input, making them more practical than hand-coding rules. In diagnosis, these models are pathophysiological models, encoding the relationships between conditions and symptoms; what diseases can occur, how they interact with the wider physiological system, and how they cause observable manifestations as signs and symptoms [69]. They describe the expected behavior of the physiological system, and discrepancies between expected and observed behavior are used to infer the presence of conditions or diseases.

This approach to diagnosis was first formalized in the late 1980s [70], and referred to as “diagnosis from first principles” due to its use of objective information about the system being diagnosed. It is also known as consistency-based diagnosis, as early implementations treated faulty systems as a logical inconsistency that is resolved by a diagnostic hypothesis that assumes some components are functioning normally and others are faulty [71]. The main drawback of these model-based approaches is that the cost of inference can be much higher than in rules-based systems [72]. For example, the query complexity of a decision tree is fixed by the tree depth, whereas a Boolean model could require solving a B-SAT problem that is NP-complete. As a result, memory-based and rules-based approaches can be used to amortize this complexity for common inputs, much in the same way that dual-process theory describes type-1 reasoning being used to amortize the more computationally expensive type-2 reasoning for common diagnostic cases.

Abductive Diagnosis

At approximately the same time that consistency-based diagnosis was developed, a

separate definition of diagnosis based on a combination of first-order logic and causal models was developed by several researchers [73, 74]. The first attempt to use causal pathophysiological models to perform diagnosis was the CASNET project of the late 1970s [75], which used a model of medical knowledge encoded as cause-effect relationships and logical conditions. Rather than basing the diagnosis on logical consistency, this approach – known as abductive diagnosis – treats diagnosis as the task of inferring the set of causes that are most likely to be responsible for the patient’s symptoms, i.e., the “best causal explanation.” Abductive inference involves reasoning from a fact (evidence, e.g., a symptom) and a causal rule (how diseases cause this symptom) to a case (diagnosis) [71]. This can be seen as a mapping from a data space of findings (e.g., symptoms and signs) to a space of diagnostic hypotheses (e.g., the presence or absence of diseases), facilitated by a causal model (Fig. 3). Unlike in rules-based diagnosis, hypotheses can interact in complex ways. For example, lower immune function can decrease a fever, while flu can increase a fever, and the two conditions can interfere and cancel out. Typically there are multiple hypotheses that can explain each piece of

evidence, and abductive inference involves enumerating the set of consistent causal hypotheses and weighting them, commonly using a heuristic measure [76]. For example, the parsimonious set covering heuristic [77] weights hypotheses based on parsimony (the number of diseases) and explanatory coverage (the number of symptoms explained), in an attempt to balance the principles of Occam’s razor and Hickam’s dictum [78].

Bayesian Diagnosis

An additional complexity of diagnostic decision-making stems from the fundamental ambiguity and uncertainty of diagnostic information. As seen in the previous section, while some model-based diagnostic systems tackle this uncertainty with fuzzy logic another approach is to extend pathophysiological models to causal Bayesian networks (BNs) [79] (see ► Chap. 11, “Artificial Intelligence for Medical Decisions” for a detailed discussion), and replace abductive heuristics with Bayesian inference. The use of Bayes’ rule in diagnosis predates even rules-based systems [80]. Compared to rules-based systems, which use forward and backward-chaining inference [81], and abductive diagnostic models, which use heuristic weighting, Bayesian models employ Bayesian inference to weight diagnostic hypotheses. This offers a more principled diagnostic weighting that incorporates uncertainty estimates, which themselves are important determinants of diagnostic decisions. Hypotheses are weighted using the posterior probability of the hypothesis given the evidence, computed via Bayes rule, and known conditional probabilities. For example, for findings f and hypothesis H the posterior is calculated as,

$$P(H|f) = \frac{P(f|H)P(H)}{P(f)} \quad (3)$$

Where the conditional probability $P(f|H)$ is the likelihood of observing a finding f given that a hypothesis H is known to be true, and $P(H)$ is the prior probability of the hypothesis being true (e.g., the prevalence of a given disease in the population)

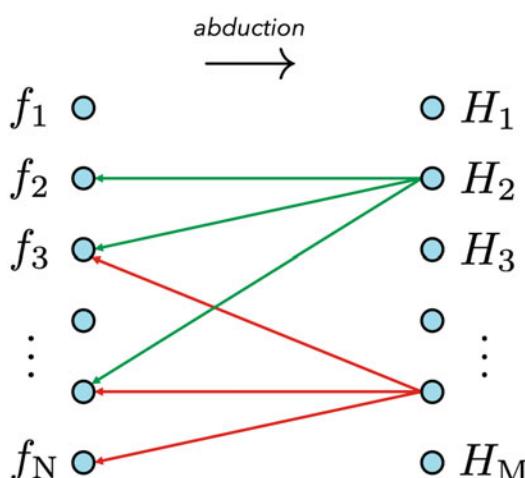


Fig. 3 Underlying diseases or conditions cause abnormal signs and symptoms (findings f). Abduction is the process of selecting diagnostic hypotheses H over these conditions, given knowledge of the findings and the causal rules that generate them

and likewise for $P(f)$. These conditional and prior probabilities are directly encoded in the causal pathophysiological model, and can be learned from multiple data sources including epidemiological data and expert knowledge (see ▶ Chap. 11, “Artificial Intelligence for Medical Decisions”).

The simplest example of a Bayesian network, still widely used for diagnosis, are naive Bayes models. These use the assumption that findings are conditionally independent given the hypothesis, providing a simplified form of the conditional probability that can be easily calculated for multiple findings,

$$P(f_1, \dots, f_N | H) = \prod_{i=1}^N P(f_i | H) \quad (4)$$

This independence assumption can be summarized as a graphical model (Fig. 4). The factorized conditional probability table is simple to learn from expert knowledge and can be easily extended to incorporate new findings, making the task of extending decision rules to larger test sets vastly simplified compared to rules-based systems. The drawback of naive Bayes classifiers is that the strong conditional independence assumption removes hypothesis interactions, resulting in an overly simplistic hypothesis weighting.

In general, BNs are generative models comprising of nodes situated in a directed acyclic graph (DAG), representing variables and their causal relations respectively. As with the naive Bayes model, BNs encode conditional independences that expedite model learning and inference (see ▶ Chap. 11, “Artificial Intelligence for Medical Decisions” for discussion). These models emerged in 1980s, most notably through the work of Judea Pearl [79]. The earliest landmark examples of model-based diagnostic systems using Bayesian networks include QMR-DT/Internist-1 [18] and later iterations of the CASNET project [55]. The first iteration of QMR-DT employed a bipartite BN, consisting of a layer of diseases (hypotheses) and symptoms (evidence). Even these simple bipartite models exhibit non-trivial hypothesis interaction such as explaining-away [83]. More

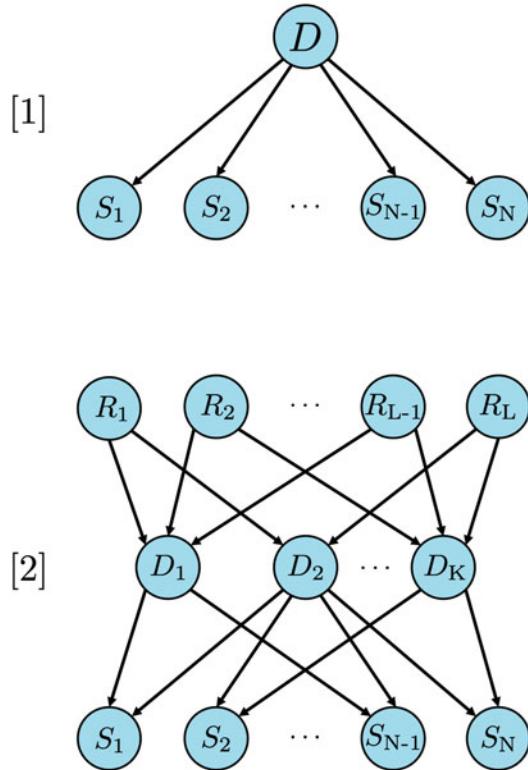


Fig. 4 [1] depicts the naive Bayes model. In this example, findings are symptoms S_1, S_2, \dots, S_N and the model incorporates a single disease hypothesis D . The joint distribution factorizes with respect to the DAG by the Markov property, $P(D, S_1, \dots, S_N) = P(D) \prod_{i=1}^N P(S_i | D)$. [2] depicts the three-layer BN extension of the CASNET network [82], including multiple disease hypotheses, a layer of symptoms (effects of diseases) and risk factors (causes of diseases)

complicated BNs (Fig. 4) with less restrictive conditional independence assumptions are more expressive and capture increasingly complex hypothesis interactions (for example, mutual exclusivity of hypotheses), allowing for more complex diagnostic decision rules at the expense of increased model learning and inference complexity [84]. The bipartite QMR-DT network was eventually replaced in [82] by three-layer BNs that incorporated an additional layer of disease risk-factors (Fig. 4), allowing for the inclusion of evidence on the causes of diseases rather than just the effects (findings). QMR-DT employs a sequential Bayesian approach to diagnosis that attempts to use Bayes rule in combination with information

gathering and hypothesis refining heuristics to mirror the active diagnostic process performed by doctors. Symptom gathering heuristics are used to rule-in/out hypotheses, gradually refining the diagnostic hypothesis set and explaining away symptoms until a final diagnostic hypothesis is returned. These early information gathering heuristics for BN models have since been superseded by more principled approaches including value of information (VOI) analysis [85, 86] and cost-utility analysis [87].

Causal Reasoning for Diagnosis

Another benefit of causal models is that they support causal inferences, establishing the effect of intervening on clinical variables (e.g., treating diseases), and counterfactuals “what-if” inferences, which can be used to quantify how well a disease hypothesis causally explains a patient’s symptoms [79, 88]. Counterfactual hypotheses are logical analogues to counterfactual thinking in human cognition, which are often used to ascribe cause-effect explanations and responsibility – for example, “the accident would not have happened if the driver had not been inebriated” is a “but-for” counterfactual hypothesis stating that driver inebriation was a necessary cause of an accident [89]. Recently, a new approach to Bayesian model-based diagnosis has been proposed that uses counterfactual inference instead of Bayes rule to weight disease hypotheses, motivated by the observation that the standard posterior weighting fails to satisfy certain common-sense postulates for diagnostic hypothesis weighting (such as assigning a weight of zero to non-causal diseases) [88]. This approach replaces the hypothesis “disease D is present,” $D = T$, with counterfactual hypotheses, for example,

$$P(S_{D=F} = F | S = T) \quad (5)$$

which is “the probability that symptom S would be absent if disease D had not been present, given that S is present,” which captures the likelihood that the presence of disease D is a necessary cause of symptom S [90]. $S_{D=F}=F$ describes the

counterfactual event that symptom S would be off, had disease D been off (counter to the fact $S = T$). By comparison, the posterior $P(D = T | S = T)$ captures only the probability that D is present, independent of whether or not it is causing the symptom S . This counterfactual approach combines the benefits of Bayesian diagnosis with the causal reasoning of abductive diagnosis and the logical reasoning of consistency-based diagnosis, and has been shown to improve the accuracy of model-based diagnostic algorithms, even outperforming human experts at differential diagnosis [88]. More details on the role of causal and counterfactual reasoning in medical decision-making are provided in ► Chap. 11, “Artificial Intelligence for Medical Decisions.”

Machine Learning for Diagnosis

Learning from Data

Machine learning methods enable the discovery of useful patterns and effective decision rules directly from data. Contrary to the rule-based methods for diagnosis, which follow a predefined series of nested queries, or model-based methods, which use causal knowledge about the world to investigate hypotheses, pattern recognition methods are akin to searching in memory for the closest match to a particular situation, in direct analogy to type 1 cognition in medical diagnosis. For example, a simple pattern recognition system could be built around a database of “patient presentations” called *exemplars*, composed of a set of clinical signs as well as their associated diagnosis. The crudest inference algorithm would then score all diseases with the average number of signs they have in common with the patient presentation and return as diagnosis the “nearest neighbor” in this medical evidence space (Fig. 5a).

While a valid approach, this simple algorithm suffers from several drawbacks that machine learning algorithms have addressed. First of all, searching an extensive database of known exemplars (specific situations), composed of millions of patient presentations, would arguably take time. A simple approach could be to produce

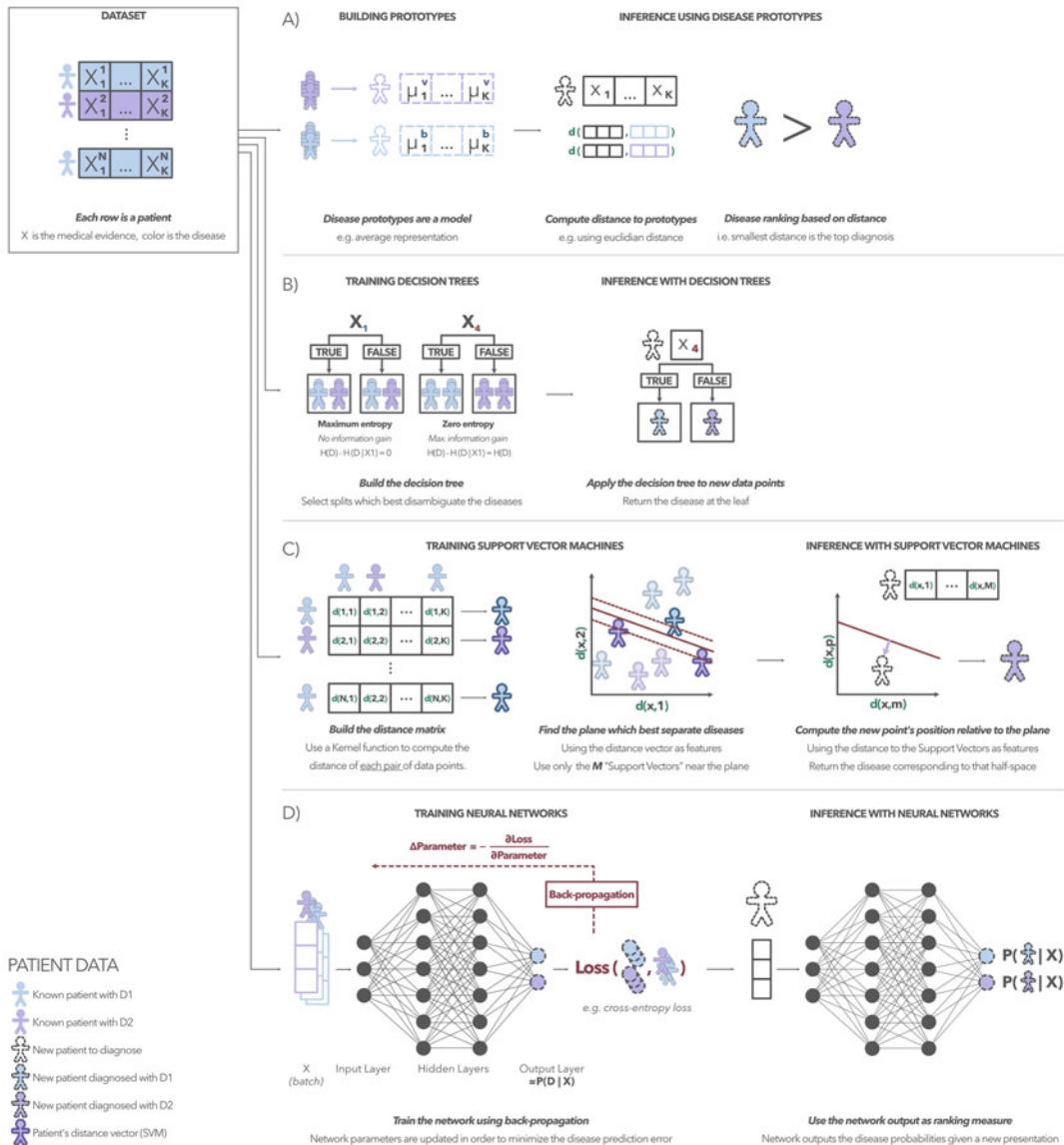


Fig. 5 Schematic description of four Machine learning approaches to diagnosis, using either (a) disease prototypes (“frames”), (b) Decision Trees, (c) Support Vector Machines, or (d) Neural Networks. The inference task

“prototypes,” an average presentation, for each specific disease. Rather than search over the space of millions of examples, the search would then be over the space of a few hundred disease prototypes and yet produce identical results. This step of “concept learning” from data, prototype building through averaging, would be an

presented here is to return the most likely hypotheses among two diseases (blue or violet). The multi-class problem of classifying more than three diseases, or returning a full hypothesis ranking is not presented here for simplicity

elementary form of machine learning. The learned “model” now being this collection of prototypes, which the counting algorithm can then use to produce diagnostic decision faster. While not learned from data, relying on such disease prototypes, called “frames,” was for example used in the CADUCEUS expert system [45].

However, by merely counting clinical signs, we would adopt many wrong assumptions. Firstly, we would assume that all signs are equally informative of a disease's presence, which is obviously wrong. The absence of a particular sign can eliminate a hypothesis, and specific "pathognomonic" signs indicate with almost complete certainty the presence of a disease. Hence, some signs should weight more in the final decision, and a better system would be able to associate a unique weight to each sign-disease pairs. Simple counting also assumes that any association of signs can be treated equally. However, clinical reasoning is full of known associations of signs (syndromes), often specific to a particular disease or class of diseases. A better system here would learn the dependencies between signs and the dependencies of groups of signs with specific diseases. This could take the form of learned weights, not only for specific sign-disease associations, but for associations between disease or classes of disease and groups of signs. Finally, counting requires the binning of continuous variables (e.g., core temperature). A better model would be able to adapt its decision smoothly over the range of the continuous variables, and ideally would even be able to rely on a derived quantity such as a polynomial function of the variable (e.g., the temperature squared). Overall, simple counting obfuscates the complex underlying processes generating the patient presentation, the dependencies between clinical signs and how strongly they are associated with their generating pathology. From a statistical standpoint, machine learning algorithms resolve those issues by learning, often implicitly, the dependencies between input features (e.g., clinical signs) and the strength of their association to the predicted target(s) (e.g., diseases).

Machine Learning as Function Approximation

As previously described, producing a measureable to rank hypotheses given a set of evidence is the answer to the *passive diagnosis* task (see Eq. 2). For each potential evidence combination, a

diagnostic function should then return a value for all hypotheses. For elementary problems, assuming the function is known, you could store all these values inside a database. However, this approach rapidly becomes intractable as the size of the evidence and hypothesis space grows. With ten binary pieces of clinical evidence (present or absent), the number of values is $2^{10} |\mathcal{H}|$, it would be $3^{10} |\mathcal{H}|$ if you account for missing evidence, and such a system could not handle continuous variables without binning. As described previously, the number of potential symptoms is in the hundreds, not counting physical signs, and para-clinical evidence. Machine learning offers the possibility to learn a function, a model, with a manageable number of parameters, which approximates the expected output. This approach relies on the assumption that there is an exploitable structure in the task which allows such approximation for example, assuming that for a small variation of the input, the output varies only slightly (smoothness) — or that the noise in the data is relatively homogeneous (homoskedasticity). As models increase in complexity, they can represent more of the data's sharp irregularities, and those assumptions loosen. Nevertheless, even for the largest known deep learning model of a trillion of parameters, the model's complexity is still several orders of magnitudes below the size of the hypothesis and input space. Those models are manageable, and fast to query — they *amortize* the complex lookup task. They are also able to generalize. They are trained to average the data and avoid *over-fitting* the noise while proposing sensible values for out-of-distribution data points (absent from the dataset), that is, performing extrapolation and interpolation.

Contrary to the natural evolutionary refinement of clear and memorable semiological descriptions seen in medicine, the decision rules learned by ML methods are often challenging to understand for humans and produce "black-box" models. The natural limit of human cognitive abilities has generally pushed semiology toward high yield combinations of signs (syndromes) and the production of simple diagnostic rules, facilitating learning and clinical practice. On the other hand, Machine learning approaches, while still limited in certain

respects (e.g., privacy, safety, explainability) [91], are best able to handle most of the task-relevant complexity present in the data. For example, all medical students can remember the characteristic description of melanoma on visual inspection as A (asymmetrical shape), B (irregular borders), C (unevenly colored), D (diameter over 6 mm), and E (evolving) [92]. However, while the simplicity and detection sensitivity of this combination of signs led to its diffusion into medical textbooks, it remains a very coarse-grained description of a highly variable skin lesion.

Recently, Deep Learning models were able to beat experienced dermatologists in diagnosing melanoma from skin lesion photographs [93]. Contrary to the simple ABCDE rule, those *computer-vision* models can learn directly from data how to best extract the information in an image to solve a particular task. Most computer-vision models used in medical research still primarily rely on the convolutional neural network architecture, which was among the first model to take full images directly as input (i.e., as X) [94, 95]. These architectures were designed to be better able to handle translational invariance, in this case, to learn useful features equally well irrespective of their absolute position in the image. We call an *inductive bias* this type of embedded property in a DL architecture. More recently, innovations in network architectures inspired from the concept of *attention* have allowed to increase the model's ability to capture complex dependencies in the data while remaining efficient [96, 97]. In the future, those approaches will probably take a more important place in the toolbox used in clinical research.

Three Supervised Methods

The field of machine learning as applied to medical diagnosis has been active for several decades. This vast family of algorithms was applied across many medical specialties covered over the next chapters. In early 2021, a PubMed query for the MeSH term “computer-assisted diagnosis” returned 82’777 results dating back as early as 1966. Eighty-six percent of those papers were

published in the first two decades of the twenty-first century, at an average rate of 3’814 publication a year. Among those, 4’084 relate to machine learning approaches (terms: Neural Network, Support Vector Machine, and Decision Trees), from 40 publications in 2000, with a peak at 987 publications in 2019. These figures constitute a loose lower bound on the actual number of research papers published on the subject, which may be published in AI conferences or journals, as well as on open platforms (e.g., ArXiv), or may not be picked up by the PubMed MeSH terms.

Among the many ML methods developed over the years, a few key algorithms are worth mentioning: decision tree algorithms, Support Vector Machines (SVM), and Deep Learning Approaches. After the enthusiasm for expert systems died down, the 1990s saw a renewed interest in data-driven, learning-based methods, beyond applied statistics. The field of ML as applied to diagnosis was then dominated by “Symbolic Learning” which aimed at discovering decision rules from data and producing models which remained understandable for clinicians. Those “decision tree” algorithms, stemming from work by Quinlan who developed the Iterative Dichotomizer 3 (ID3), work in a fundamentally similar way to “prototype building.” In our case, given a set of “exemplars” (the dataset of evidence associated with a disease) we want to produce the most discriminative prototype for a disease. To do so, the algorithm creates a sequence of branching conditions where one piece of evidence is “split” at a specific value. This value is picked to maximally reduce the entropy, or uncertainty, associated with the diseases of the remaining exemplars on both sides of the fork. In other words, the split is chosen so that it gives the most information about the exemplars’ disease. The optimal branching decision would then group all the cases with disease A on one side, and all the cases of disease B on the other (Fig. 5b). This algorithm and its derivatives (e.g., random forest, extreme gradient boosting) are still extremely prevalent and are often a key part of the pipelines which win data science competitions [98]. For example, they have been used efficiently for diagnosis in cardiology [99], gastroenterology [100],

endocrinology [101], and can even beat SVMs and DL methods on those tasks [102].

While decision trees algorithms are powerful and produce transparent models, they are limited to features (here, the evidence variables) in their raw form. They cannot learn to compose features into more complex variables, for example, from the evidence variable t for “core temperature,” they would not be able to use the variable t^2 to make a decision. In regression trees, for example, we say that it produces models that are “piecewise-linear” in the features. Similarly to polynomial models [103], given certain assumptions, both SVM [104] and DL [105] methods are “universal approximators,” that is they are able to produce models approximating any function up to arbitrary small error. While considerably different, both methods produce complex models, which have decision boundaries that are non-linear in the features, offering gains in model expressivity, i.e., the computation it can perform [106], at the cost of transparency and explainability.

In broad terms, modern SVMs [107] rely on a general notion of distance between data points, performing a high-dimensional triangulation of sorts. In the case of diagnosis, it would compute the distance between each pair of patients’ presentations in the dataset using a distance function called the *kernel function*. Once all distances have been computed, they would become the algorithm’s new features (see Fig. 5c). The SVM algorithm would then segment this space of distances in two regions by finding a hyperplane which maximally separates the labels (e.g., diseases) associated with the data points (e.g., patients). During inference, given a new point (e.g., patient presentation), the algorithm would first compute the pairwise distances with the rest of the dataset, and then *classify* (e.g., diagnose) the new point with the label of the half-space it belongs to. However, keeping in memory the whole dataset to compute all those distances would not be efficient, or even feasible for large datasets. A key insight of the SVM algorithm is that only a few points, called “the support vectors,” located near the hyperplane are necessary to define it – a property exploited to produce smaller models and

fast inference. In 2017, SVMs accounted for around 40% of the ML papers in healthcare [108] and have been used extensively for diagnosis from structured data. For example, they have been used for diagnosing diabetes [109, 110], or breast cancer [111, 112]. SVMs have also been used to classify medical images, such as mammograms [113], OCT [114], or slit lamp lens images [115]. However, contrary to the more flexible and scalable DL approaches, SVMs cannot handle images directly and requires a pre-processing step to extract relevant image features [116].

As described in the previous chapter, DL approaches can handle any input and produce any output while relying on the same learning algorithm (see backpropagation, Fig. 5d) [117]. Although feature engineering can be helpful, those architectures can be trained “end to end,” without requiring a pre-processing step. In 2017, DL approaches accounted for a third of ML papers for healthcare available on PubMed [108]. From medical imagery [118], Deep Learning architectures have proven extremely efficient, producing super-human results on complex tasks, and versatile, with the same models able to solve tasks on different imagery modalities [119]. They have allowed to directly solve diagnostic tasks, notably reaching human or super-humans levels in ophthalmology, dermatology, gastroenterology, as well as cancer detection across many modalities (see [120] for a recent review). But DL has also facilitated diagnosis by solving for supporting tasks, such as image segmentation [121], feature extraction [122], staging [123], image enhancement [124], and registration [125].

Other Machine Learning Formalisms for Diagnosis

As described in the previous chapter, the passive diagnostic task is fundamentally a classification task, hence a “Supervised Learning” task. Indeed, it has an identified target to predict (i.e., the hypothesis ranking measure) from a distinct set of input variables (i.e., clinical evidence). However, it is worth mentioning that unsupervised learning approaches have a long

history in computer-assisted diagnosis for solving supporting tasks [126]. Methods such as dimensionality reduction (e.g., PCA, ICA, SVD), clustering (e.g., K-means neighbors), or more recently Auto-Encoders [127], Generative Adversarial Networks [128], or contrastive methods [129] have been used in the prepossessing steps of ML pipelines to perform feature extraction [130], perform data-augmentation [131], but also for improving data visualization and understanding [132], query medical databases for similar known cases [133], or enhance medical images [134].

Reinforcement Learning has been used for diagnosis and support tasks as well. RL is particularly suited to solve the combinatorial optimization task of learning to gather information during active diagnosis [135–137]. Rather than live interaction with real patients, the RL agent relies on an offline dataset of presentations associated with target label (e.g., disease) to learn whether to ask for more information or stop the process and propose a diagnosis. To our knowledge, RL is the only learning-based approach used to tackle active diagnosis so far, but sequence-to-sequence models, such as the novel transformers [96], should also be able to solve that task in the future.

The Importance of Data

A key blocker in the application of ML in healthcare is the relative scarcity of publicly available datasets. ML algorithms require data to learn models, and these models can only be as good as the quality of the training data allows. Biases and low signal to noise ratio in the data will directly impact the models, above and beyond the choice of learning algorithm used to train those models. Data is the critical resource of ML, and the field thrives on quality datasets to build robust benchmarks that allow for fair comparisons and to measure the field’s progress. However, despite the exabytes of health data produced each day, for obvious reasons of privacy there is still only a few datasets made publicly available (e.g., MIMIC [138], Deep Lesion [139], OASIS [140], openfMRI [141], CT Medical Image [142], EyePACS-1 [143], STARE [144]). In recent

years, this problem has been tackled head-on by several countries and universities which have increased their effort to making anonymized healthcare data publicly available to facilitate research [145, 146].

Outlook

In this chapter, we have presented an overview of diagnosis both as an active task, explicitly involving interaction with the patient and data gathering, and as a passive classification task, and have explored how different AI methods are complementary to these different definitions. As diagnostic AI is primarily employed for decision support and often ultimately ends up in doctors’ hands, we have aimed to present diagnosis within the context of research into human diagnostic reasoning and the cognitive causes of diagnostic errors. These research areas can shine a light on the issues facing doctors in clinical practice, and we have argued that these factors can guide the development and deployment of effective diagnostic decision support systems. We then reviewed each of the three main approaches to AI-assisted diagnosis, detailing how they capture the diagnostic process and which diagnostic tasks they are best suited for. While these methods were borne out of different times, and often fundamentally different approaches to AI, each of them is an important tool that can be used for building diagnostic AI. Diagnosis is a special instance of clinical decision-making, and in the previous ▶ Chap. 11, “Artificial Intelligence for Medical Decisions” we outlined the importance of explainability, safety, and fairness in clinical decision algorithms. These issues are equally important in diagnostic algorithms, and we refer the reader to these sections. One factor that we have left out of this chapter is the importance of design and implementation in diagnostic decision support, which has arguably been one of the major constraints preventing the widespread adoption of AI solutions. This is a large topic and is heavily dependent on the specifics of hospitals, healthcare systems and clinical practice, which go beyond the scope of this chapter. For this we refer the reader to follow review articles for diagnostic decision support systems [147, 148] and medical AI in general [149, 108, 150].

References

1. Newman-Toker DE, Pronovost PJ. Diagnostic error—the next frontier for patient safety. *JAMA*. 2009;301(10):1060–2.
2. Gruber ML. The incidence of diagnostic error in medicine. *BMJ Qual Saf*. 2013;22(Suppl 2):ii21–7.
3. Singh H, Schiff GD, Gruber ML, Onakpoya I, Thompson MJ. The global burden of diagnostic errors in primary care. *BMJ Qual Saf*. 2017;26(6):484–94.
4. Newman-Toker DE, McDonald KM, Meltzer DO. How much diagnostic safety can we afford, and how should we decide? A health economics perspective. *BMJ Qual Saf*. 2013;22(Suppl 2):ii11–20.
5. Diagnostic Errors: Technical Series on Safer Primary Care. Geneva: World Health Organization; 2016. Licence: CC BY-NC-SA 3.0 IGO.
6. Sadegh-Zadeh K. Fuzzy logic. In: *Handbook of analytic philosophy of medicine*. Netherlands: Springer; 2015. p. 1055–110.
7. Sampath M, Lafortune S, Teneketzi D. Active diagnosis of discrete-event systems. *IEEE Trans Autom Control*. 1998;43(7):908–29.
8. Peirce CS. Philosophical writings of Peirce (J. Buchler, ed). Vol 217. New York: Dover. 1955.
9. Ramoni M, Stefanelli M, Magnani L, Barosi G. An epistemological framework for medical knowledge-based systems. *IEEE Trans Syst Man Cybern*. 1992;22(6):1361–75.
10. Patel VL, Arocha JF, Jiajie Z. Thinking and reasoning in medicine. In: *The Oxford handbook of thinking and reasoning*. Oxford University Press, 2012. p. 1–34.
11. Simmons B. Clinical reasoning: concept analysis. *J Adv Nurs*. 2010;66(5):1151–8.
12. Pelaccia T, Tardif J, Triby E, Charlin B. An analysis of clinical reasoning through a recent and comprehensive approach: the dual-process theory. *Med Educ Online*. 2011;16(1):5890.
13. Evans JSBT, Stanovich KE. Dual-process theories of higher cognition: advancing the debate. *Perspect Psychol Sci*. 2013;8(3):223–41.
14. Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis. *Science*. 1959;130(3366):9–21.
15. Sloman SA. The empirical case for two systems of reasoning. *Psychol Bull*. 1996;119(1):3.
16. Mao Q, Jay M, Hoffman JL, Calvert J, Barton C, Shimabukuro D, Shieh L, Chettipally U, Fletcher G, Kerem Y, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. *BMJ Open*. 2018;8(1):e017833,2018.
17. Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, Kalloo A, Hassen ABH, Thomas L, Enk A, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol*. 2018;29(8):1836–42.
18. Shwe MA, Middleton B, Heckerman DE, Henrion M, Horvitz EJ, Lehmann HP, Cooper GF. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. *Methods Inf Med*. 1991;30(4):241–55.
19. Bakator M, Radosav D. Deep learning and medical diagnosis: a review of literature. *Multimod Technol Interact*. 2018;2(3):47.
20. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Disc*. 2019;9(4):e1312.
21. Marcus G. Deep learning: a critical appraisal. *arXiv preprint arXiv:1801.00631*. 2018.
22. Geirhos R, Jacobsen J-H, Michaelis C, Zemel R, Brendel W, Bethge M, Wichmann FA. Shortcut learning in deep neural networks. *Nat Mach Intell*. 2020;2(11):665–73.
23. Badgley MA, Zech JR, Oakden-Rayner L, Glicksberg BS, Liu M, Gale W, McConnell MV, Percha B, Snyder TM, Dudley JT. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digit Med*. 2019;2(1):1–10.
24. DeGrave AJ, Janizek JD, Lee S-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *medRxiv*. 2020.
25. Berner ES. Clinical decision support systems, vol. 233. New York, NY, USA: Springer; 2007.
26. Castaneda C, Nalley K, Mannion C, Bhattacharyya P, Blake P, Pecora A, Goy A, Suh KS. Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. *J Clin Bioinform*. 2015;5(1):1–16.
27. Garg AX, Adhikari NKJ, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, Sam J, Haynes RB. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *JAMA*. 2005;293(10):1223–38.
28. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. *Acad Med*. 2003;78(8):775–80.
29. Bruno MA, Walker EA, AbuJudeh HH. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. *Radiographics*. 2015;35(6):1668–76.
30. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer*. 2018;18(8):500–10.
31. Busby LP, Courtier JL, Glastonbury CM. Bias in radiology: the how and why of misses and misinterpretations. *Radiographics*. 2018;38(1):236–47.
32. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, Back T, Chesus M, Corrado GS, Darzi A, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577(7788):89–94.
33. Gruber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. *Arch Intern Med*. 2005;165(13):1493–9.

34. Crowley RS, Legowski E, Medvedeva O, Reitmeyer K, Tseytin E, Castine M, Jukic D, Mello-Thoms C. Automated detection of heuristics and biases among pathologists in a computer-based system. *Adv Health Sci Educ.* 2013;18(3):343–63.
35. Rotmensch M, Halpern Y, Tlimat A, Horng S, Sontag D. Learning a health knowledge graph from electronic medical records. *Sci Rep.* 2017;7(1):1–11.
36. Shortliffe EH. Mycin: a knowledge-based computer program applied to infectious diseases. In: Proceedings of the Annual Symposium on Computer Application in Medical Care. American Medical Informatics Association; 1977. p. 66.
37. Aikins JS, Kunz JC, Shortliffe EH, Fallat RJ. Puff: an expert system for interpretation of pulmonary function data. *Comput Biomed Res.* 1983;16(3):199–208.
38. Kingsland LC, Lindberg DAB, Sharp GC. AI/RHEUM. *J Med Syst.* 1983;7(3):221–7.
39. Adlassnig K-P, Kolarz G, Scheithauer W, Effenberger H, Grabner G. CADIG: approaches to computer-assisted medical diagnosis. *Comput Biol Med.* 1985;15(5):315–35.
40. Zadeh LA. Information and control. *Fuzzy Sets.* 1965;8(3):338–53.
41. Fieschi M, Joubert M, Fieschi D, Roux M. SPHINX – a system for computer-aided diagnosis. *Methods Inf Med.* 1982;21(03):143–8.
42. Godo LL, de Mántaras RL, Sierra C, Verdaguer A. Managing linguistically expressed uncertainty in milord application to medical diagnosis. *AI Commun.* 1988;1(1):14–31.
43. Lekkas S, Mikhailov L. Evolving fuzzy medical diagnosis of Pima Indians diabetes and of dermatological diseases. *Artif Intell Med.* 2010;50(2):117–26.
44. Kour H, Manhas J, Sharma V. Usage and implementation of neuro-fuzzy systems for classification and prediction in the diagnosis of different types of medical disorders: a decade review. *Artif Intell Rev.* 2020;53(7):4651–706.
45. Myers JD, Pople HE, Miller RA. Caduceus: a computerized diagnostic consultation system in internal medicine. In: Proceedings of the Annual Symposium on Computer Application in Medical Care. American Medical Informatics Association; 1982. p. 44.
46. Köhler S, Schulz MH, Krawitz P, Bauer S, Dölkens S, Ott CE, Mundlos C, Horn D, Mundlos S, Robinson PN. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet.* 2009;85(4):457–64.
47. Gounot VB, Donfack V, Lasbleiz J, Bourde A, Duvaufier R. Creating an ontology driven rules base for an expert system for medical diagnosis. *Stud Health Technol Inform.* 2011;169:714–8.
48. Kazemi SM, Poole D. Simple embedding for link prediction in knowledge graphs. arXiv preprint arXiv:1802.04868. 2018.
49. Lukovnikov D, Fischer A, Lehmann J, Auer S. Neural network-based question answering over knowledge graphs on word and character level. In: Proceedings of the 26th International Conference on World Wide Web. 2017. p. 1211–20.
50. Algergawy A, Cheatham M, Faria D, Ferrara A, Fundulaki I, Harrow I, Hertling S, Jiménez-Ruiz E, Karam N, Khiat A, et al. Results of the ontology alignment evaluation initiative 2018. In: 13th International Workshop on Ontology Matching co-located with the 17th ISWC (OM 2018), vol. 2288. 2018. p. 76–116.
51. World Health Organization. International statistical classification of diseases and related health problems: tabular list, vol. 1. Geneva, Switzerland: World Health Organization; 2004.
52. SNOMED International. SNOMED-CT.
53. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 2004;32(suppl 1):D267–70.
54. Lee D, de Keizer N, Lau F, Cornet R. Literature review of SNOMED CT use. *J Am Med Inform Assoc.* 2014;21(e1):e11–9.
55. Weiss SM, Kulikowski CA, Safir A. A model-based consultation system for the long-term management of glaucoma. In: IJCAI, vol. 5. 1977. p. 826–32.
56. Lauritzen SL, Spiegelhalter DJ. Local computations with probabilities on graphical structures and their application to expert systems. *J R Stat Soc Ser B Methodol.* 1988;50(2):157–94.
57. Miller RA, McNeil MA, Challinor SM, Masarie FE Jr, Myers JD. The INTERNIST-1/Quick Medical Reference project status report. *West J Med.* 1986;145(6):816.
58. INSERM. Orphanet.
59. Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, Black GCM, Brown DL, Brudno M, Campbell J, et al. The human phenotype ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.* 2014;42(D1):D966–74.
60. Pinchin V. I'm feeling yucky :(searching for symptoms on google. The Keyword, 2016.
61. Turki H, Shafee T, Taieb MAH, Aouicha MB, Vrandecić D, Das D, Hamdi H. Wikidata: a largescale collaborative ontological medical database. *J Biomed Inform.* 2019;99:103292.
62. Abbasi J. Shantanu Nundy, MD: the human diagnosis project. *JAMA.* 2018;319(4):329–31.
63. De Dombal FT, Leaper DJ, Horrocks JC, Staniland JR, McCann AP. Human and computer-aided diagnosis of abdominal pain: further report with emphasis on performance of clinicians. *Br Med J.* 1974;1 (5904):376–80.
64. Lucas PJF. Symbolic diagnosis and its formalisation. *Knowl Eng Rev.* 1997;12(2):109–46.
65. Partridge D. The scope and limitations of first generation expert systems. *Futur Gener Comput Syst.* 1987;3(1):1–10.
66. Van De Riet RP. Problems with expert systems? *Futur Gener Comput Syst.* 1987;3(1):11–6.
67. Davis R. Expert systems: where are we? And where do we go from here? *AI Mag.* 1982;3(2):3–3.
68. Mozetič I. Model-based diagnosis: an overview. In: Mřík V, Štěpánková O, Trappi R, editors. Advanced topics in artificial intelligence. Berlin/Heidelberg: Springer; 1992. p. 419–30.

69. Bylander T. Some causal models are deeper than others. *Artif Intell Med.* 1990;2(3):123–8.
70. Reiter R. A theory of diagnosis from first principles. *Artif Intell.* 1987;32(1):57–95.
71. Poole D. Normality and faults in logic-based diagnosis. In: IJCAI, vol. 89. Citeseer; 1989. p. 1304–10.
72. Eiter T, Gottlob G. The complexity of logic-based abduction. *J ACM.* 1995;42(1):3–42.
73. Cox PT, Pietrzykowski T. General diagnosis by abductive inference. In: SLP, vol. 183. 1987. p. 189.
74. Poole D, Goebel R, Aleliunas R. Theorist: a logical reasoning system for defaults and diagnosis. In: The knowledge frontier. Springer-Verlag, Berlin; 1987. p. 331–52.
75. Weiss SM, Kulikowski CA, Amarel S, Safir A. A model-based method for computer-aided medical decision-making. *Artif Intell.* 1978;11(1–2):145–72.
76. Finin T, Morris G. Abductive reasoning in multiple fault diagnosis. *Artif Intell Rev.* 1989;3(2):129–58.
77. Reggia JA, Nau DS, Wang PY. A formal model of diagnostic inference. I. Problem formulation and decomposition. *Inf Sci.* 1985;37(13):227–56.
78. Mani N, Slevin N, Hudson A. What three wise men have to say about diagnosis. *BMJ.* 2011;343:d7769.
79. Pearl J. Causality (2nd ed.). Cambridge: Cambridge University Press; 2009.
80. Gorry GA, Barnett GO. Experience with a model of sequential diagnosis. *Comput Biomed Res.* 1968;1(5):490–507.
81. Musen MA, Middleton B, Greenes RA. Clinical decision support systems. In: Biomedical informatics. Springer, New York; 2014. p. 643–74.
82. Pradhan M, Provan G, Middleton B, Henrion M. Knowledge engineering for large belief networks. In: Uncertainty Proceedings 1994. Elsevier; 1994. p. 484–90.
83. Wellman MP, Henrion M. Explaining ‘explaining away’. *IEEE Trans Pattern Anal Mach Intell.* 1993;15(3):287–92.
84. Pourret O, Naïm P, Marcot B. Bayesian networks: a practical guide to applications. Wiley, West Sussex, England; 2008.
85. Yokota F, Thompson KM. Value of information literature analysis: a review of applications in health risk management. *Med Decis Mak.* 2004;24(3):287–98.
86. Buchard A, Baker A, Gourgoulias K, Navarro A, Perov Y, Zwiesele M, Johri S. Tuning semantic consistency of active medical diagnosis: a walk on the semantic simplex. In: Frontier of AI-Assisted Care (FAC) Scientific Symposium. 2019.
87. Shachter RD. Evaluating influence diagrams. *Oper Res.* 1986;34(6):871–82.
88. Richens JG, Lee CM, Johri S. Improving the accuracy of medical diagnosis with causal machine learning. *Nat Commun.* 2020;11(1):1–9.
89. Halpern JY. Actual causality. Cambridge, MA: MIT Press; 2016.
90. Pearl J. Probabilities of causation: three counterfactual interpretations and their identification. *Synthese.* 1999;121(1):93–149.
91. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, Ferrero E, Agapow P-M, Zietz M, Hoffman MM, et al. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface.* 2018;15(141):20170387.
92. Abbasi NR, Shaw HM, Rigel DS, Friedman RJ, McCarthy WH, Osman I, Kopf AW, Polsky D. Early diagnosis of cutaneous melanoma: revisiting the ABCD criteria. *JAMA.* 2004;292(22):2771–6.
93. Goyal M, Knackstedt T, Yan S, Hassanpour S. Artificial intelligence-based image classification for diagnosis of skin cancer: challenges and opportunities. *Comput Biol Med.* 2020;127:104065.
94. Fukushima K. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics.* Springer-Verlag, Berlin; 1980;36:193–202.
95. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* 1989;1(4):541–51.
96. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. In: Advances in neural information processing systems; 2017. p. 5998–6008. Cambridge, MA: MIT Press.
97. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jegou H. Training data-efficient image transformers & distillation through attention. arXiv preprint arXiv:2012.12877. 2020.
98. Goldblum A. What algorithms are most successful on Kaggle? 2016. Available at: <https://www.kaggle.com/antgoldblum/what-algorithms-are-most-successful-on-kaggle>. (Accessed: 9th August 2021).
99. Pavlopoulos SA, Stasis ACH, Loukis EN. A decision tree-based method for the differential diagnosis of aortic stenosis from mitral regurgitation using heart sounds. *Biomed Eng Online.* 2004;3(1):1–15.
100. Zorman M, Eich H-P, Kokol P, Ohmann C. Comparison of three databases with a decision tree approach in the medical field of acute appendicitis. *Stud Health Technol Inform.* 2001;84 (Pt 2):1414–8.
101. Habibi S, Ahmadi M, Alizadeh S. Type 2 diabetes mellitus screening and risk factors using decision tree: results of data mining. *Global J Health Sci.* 2015;7(5):304.
102. Lee H-C, Yoon H-K, Nam K, Cho YJ, Kim TK, Kim WH, Bahk J-H. Derivation and validation of machine learning approaches to predict acute kidney injury after cardiac surgery. *J Clin Med.* 2018;7(10):322.
103. Stone MH. The generalized weierstrass approximation theorem. *Math Mag.* 1948;21(5):237–54.
104. Hammer B, Gersmann K. A note on the universal approximation capability of support vector machines. *Neural Process Lett.* 2003;17(1):43–53.
105. Cybenko G. Approximation by superpositions of a sigmoidal function. *Math Control Signals Syst.* 1989;2(4):303–14.
106. Raghu M, Poole B, Kleinberg J, Ganguli S, Sohl-Dickstein J. On the expressive power of deep neural

- networks. In: International Conference on Machine Learning. PMLR; 2017. p. 2847–54.
107. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. 1992. p. 144–52.
 108. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, Wang Y. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol*. 2017;2(4):230.
 109. Stoean R, Stocean C, Preuss M, El-Darzi E, Dumitrescu D. Evolutionary support vector machines for diabetes mellitus diagnosis. In: 2006 3rd International IEEE Conference Intelligent Systems. IEEE; 2006. p. 182–7.
 110. Barakat N, Bradley AP, Barakat MNH. Intelligible support vector machines for diagnosis of diabetes mellitus. *IEEE Trans Inf Technol Biomed*. 2010;14(4):1114–20.
 111. Bennett KP, Blue JA. A support vector machine approach to decision trees. In: 1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No. 98CH36227), vol. 3. IEEE; 1998. p. 2396–401.
 112. Polat K, Güneş S. Breast cancer diagnosis using least square support vector machine. *Digital Signal Process*. 2007;17(4):694–701.
 113. Sharma S, Khanna P. Computer-aided diagnosis of malignant mammograms using Zernike moments and SVM. *J Digit Imaging*. 2015;28(1):77–90.
 114. Srinivasan PP, Kim LA, Mettu PS, Cousins SW, Comer GM, Izatt JA, Farsiu S. Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images. *Biomed Opt Express*. 2014;5(10):3568–77.
 115. Li H, Lim JH, Liu J, Mitchell P, Tan AG, Wang JJ, Wong TY. A computer-aided diagnosis system of nuclear cataract. *IEEE Trans Biomed Eng*. 2010;57(7):1690–8.
 116. Ergin S, Kilinc O. A new feature extraction framework based on wavelets for breast cancer diagnosis. *Comput Biol Med*. 2014;51:171–82.
 117. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature*. 1986;323(6088):533–6.
 118. Poudel RPK, Lamata P, Montana G. Recurrent fully convolutional neural networks for multi-slice MRI cardiac segmentation. In: Reconstruction, segmentation, and analysis of medical images. 1st International Workshops on Reconstruction and Analysis of Moving Body Organs, RAMBO 2016 and 1st International Workshops on Whole-Heart and Great Vessel Segmentation from 3D Cardiovascular MRI in Congenital Heart Disease, HVSMR 2016 (2016). p. 83–94. Springer International; 2016. p. 83–94.
 119. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, McKeown A, Yang G, Wu X, Yan F, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*. 2018;172(5):1122–31.
 120. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G, Thrun S, Dean J. A guide to deep learning in healthcare. *Nat Med*. 2019;25(1):24–9.
 121. Kamnitsas K, Ledig C, Newcombe VFJ, Simpson JP, Kane AD, Menon DK, Rueckert D, Glocker B. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal*. 2017;36:61–78.
 122. Schlegl T, Waldstein SM, Bogunovic H, Endstraßer F, Sadeghipour A, Philip A-M, Podkowinski D, Gerendas BS, Langs G, Schmidt-Erfurth U. Fully automated detection and quantification of macular UID in OCT using deep learning. *Ophthalmology*. 2018;125(4):549–58.
 123. Wolterink JM, Leiner T, Viergever MA, Işgum I. Automatic coronary calcium scoring in cardiac CT angiography using convolutional neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2015. p. 589–96.
 124. Pham C-H, Ducournau A, Fablet R, Rousseau F. Brain MRI super-resolution using deep 3D convolutional networks. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). IEEE; 2017. p. 197–200.
 125. Boveiri HR, Khayami R, Javidan R, Mehdizadeh A. Medical image registration using deep neural networks: a comprehensive review. *Comput Electr Eng*. 2020;87:106767.
 126. Raza K, Singh NK. A tour of unsupervised deep learning for medical image analysis. *arXiv preprint arXiv:1812.07715*. 2018.
 127. Hinton GE, Zemel RS. Autoencoders, minimum description length, and Helmholtz free energy. *Adv Neural Inf Proces Syst*. 1994;6:3–10.
 128. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, DavidWarde-Farley SO, Courville A, Bengio Y. Generative adversarial nets. In: Advances in neural information processing systems. 2014. p. 2672–80. Cambridge, MA: MIT Press.
 129. Chaitanya K, Erdil E, Karani N, Konukoglu E. Contrastive learning of global and local features for medical image segmentation with limited annotations. *arXiv preprint arXiv:2006.10511*. 2020.
 130. Singh G, Samavedham L. Unsupervised learning based feature extraction for differential diagnosis of neurodegenerative diseases: a case study on early-stage diagnosis of Parkinson disease. *J Neurosci Methods*. 2015;256:30–40.
 131. Zunair H, Hamza AB. Melanoma detection using adversarial training and deep transfer learning. *Phys Med Biol*. 2020;65(13):135005.
 132. Guo S, Xu K, Zhao R, Gotz D, Zha H, Cao N. EventThread: visual summarization and stage analysis of event sequence data. *IEEE Trans Vis Comput Graph*. 2017;24(1):56–65.
 133. Deepak S, Ameer PM. Retrieval of brain MRI with tumor using contrastive loss based similarity on

- GoogLeNet encodings. *Comput Biol Med.* 2020;125:103993.
134. Armanious K, Jiang C, Fischer M, Küstner T, Hepp T, Nikolaou K, Gatidis S, Yang B. Medgan: medical image translation using GANs. *Comput Med Imaging Graph.* 2020;79:101684.
135. Tang K-F. Inquire and diagnose: neural symptom checking ensemble using deep reinforcement learning. In: 29th Conference on Neural Information Processing Systems (NIPS 2016); 2016. p. 1–9.
136. Stensmo M, Sejnowski TJ. Automated medical diagnosis based on decision theory and learning from cases. *World congress on neural Net-works.* 1996;1227–1231.
137. Buchard A, Bouvier B, Prando G, Beard R, Livieratos M, Busbridge D, Thompson D, Richens J, Zhang Y, Baker A, et al. Learning medical triage from clinicians using deep q-learning. *arXiv preprint arXiv:2003.12828.* 2020.
138. Johnson AEW, Pollard TJ, Shen L, Li-Wei HL, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. Mimic-III, a freely accessible critical care database. *Sci Data.* 2016;3(1):1–9.
139. Yan K, Wang X, Lu L, Summers RM. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *J Med Imaging.* 2018;5(3):036501.
140. Marcus DS, Wang TH, Parker J, Csermansky JG, Morris JC, Buckner RL. Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *J Cogn Neurosci.* 2007;19(9):1498–507.
141. Poldrack RA, Barch DM, Mitchell J, Wager T, Wagner AD, Devlin JT, Cumba C, Koyejo O, Milham M. Toward open sharing of task-based fMRI data: the OpenfMRI project. *Front Neuroinform.* 2013;7:12.
142. Albertina B, Watson M, Holback C, Jarosz R, Kirk S, Lee Y, Lemmerman J. Radiology Data from The Cancer Genome Atlas Lung Adenocarcinoma [TCGA-LUAD] collection. The Cancer Imaging Archive. 2016. <https://doi.org/10.7937/K9/TCIA.2016.JGNIHEP5> Available at: <https://wiki.cancerimagingarchive.net/display/Public/TCGA-LUAD> (Accessed: 9th August 2021).
143. Cuadros J, Bresnick G. EyePACS: an adaptable telemedicine system for diabetic retinopathy screening. *J Diabetes Sci Technol.* 2009;3(3):509–16.
144. Hoover AD, Kouznetsova V, Goldbaum M. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Trans Med Imaging.* 2000;19(3):203–10.
145. Cuggia M, Combes S. The French health data hub and the German medical informatics initiatives: two national projects to promote data sharing in healthcare. *Yearb Med Inform.* 2019;28(1):195.
146. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong ICK, Rijnbeek PR, et al. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform.* 2015;216:574.
147. Miller RA. Computer-assisted diagnostic decision support: history, challenges, and possible paths forward. *Adv Health Sci Educ.* 2009;14(1):89–106.
148. Bright TJ, Wong A, Dhurjati R, Bristow E, Bastian L, Coeytaux RR, Samsa G, Hasselblad V, Williams JW, Musty MD, et al. Effect of clinical decision-support systems: a systematic review. *Ann Intern Med.* 2012;157(1):29–43.
149. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 2019;17(1):1–9.
150. Shortliffe EH, Sepulveda MJ. Clinical decision support in the era of artificial intelligence. *JAMA.* 2018;320(21):2199–200.



AIM and the History of Medicine

13

Kadircan H. Keskinbora

Contents

Introduction to the History of Cognitive Science and Intelligence	204
Mechanical and Biological Automatons	204
The Golem	205
The Ars Magna	205
The Concept of Symbolic Languages	205
Hurufism	205
The Calculator	205
The Technological Myth Persists	206
AI is on its Way	206
Conceptual Revolutions	207
World, Meet the Personal Computer	208
The Big Question: Can Computers Think?	209
Earliest Steps towards Computerized Medicine	209
Collaboration Between Medical Scientists and AI Researchers	210
Examples of the Use of AI in Medicine	210
Artificial and Biological Life	211
Conclusion	211
References	213

Abstract

Supplementary Information: The online version of this chapter (https://doi.org/10.1007/978-3-030-64573-1_305) contains supplementary material, which is available to authorized users.

K. H. Keskinbora (✉)
School of Medicine, Bahcesehir University, Istanbul,
Turkey

While Artificial Intelligence is thought to be a twenty-first-century development, it can be traced back to earlier times. Throughout history, philosophers have been influenced by some versions of the idea of other consciousnesses. Plato suggested that all material objects have absolute forms in his *The Dialogues*, in

the form of a series of conversations with his teacher, Socrates. This was clearly an attempt to discuss the first efforts in the representation of information. Furthermore, these early musings by Plato also display structured efforts to solve problems.

A project designed for one purpose can yield outcomes other than what it originally intended. This section will track the historical traces of Artificial Intelligence and mention some uses of artificial intelligence in medicine. Research on intelligence could lead to the creation of machines that can outdo the thinking capabilities of humans in fields like “calculating, decision making, memory, learning, creating and planning,” which are all areas that require expertise. There could also be concerning consequences.

Keywords

Artificial intelligence (AI) · History of Medicine · Symbolic languages · Numbers · figures · Hurufism · Myth · Automaton · Computerized medicine · Thinking machine

Introduction to the History of Cognitive Science and Intelligence

In her book named, *Building the Second Mind: 1956 and the Origins of Artificial Intelligence Computing*, Rebecca E. Skinner [1] explores the basic concepts of artificial intelligence (AI) in ancient tales and philosophical works.

The first conceptual examples of AI came up in Athens in the sixth century B.C. Ancient Greek thinkers were less interested in the meaning of numbers alone than in expressing intelligence and rational ideas using numbers. They produced the first expressions of the reality of abstract ideas, the first geometrical proof for the logical forms for argumentation, and put forth a primitive version of the theory of mind. Plato (429–347 BC) [2] suggested that that all material objects have an absolute form through a series of discussions with his mentor Socrates in *The Dialogues*. This was the first effort for the representation of

knowledge or a clear way to discuss ideas of various sorts. Furthermore, these early proofs found in *The Dialogues* seem to have been Plato’s endeavors to solve formulated problems. Plato’s student Aristoteles (384–322 BC) [3] also carried out systematic research into questions in the field of natural sciences. This is reminiscent of a prehistoric explanation for the initial condition and the targeted condition or approaches a search as it is called in AI. In fact, it approximates the search in all kinds of problematic fields. The concept of protocol was mentioned in the mid-twentieth century by philosopher George Polya as “an Ancient Greek intuitive approach” defined as “an adjective meaning to serve discovery.” [3] The philosopher Euclid’s dream of “representing geometric figures in space” was set up as a part of the field of geometry [3].

Mechanical and Biological Automatons

Automatons are relatively self-operating machines designed to automatically follow encoded instructions. Building automatons in the sense of recreating a human has been a recurring theme explored in many essays, novels, and fictional writings throughout history. The word automaton has Latin roots, based on an Ancient Greek machine that had the ability to move on its own. The so-called recreation of the human body and intelligence was defined in the Babylonian Epic of Gilgamesh and the Hebrew human conceptualization that God created from earth. Mechanical automatons and prosthetics with automaton-like features abound in all ancient mythologies, particularly in Greek mythology, which took on many features from preceding civilizations. Legend has it that Hephaestus, the Olympian God of fire and metalworking incorporated automata into Achilles’s shield. The Greek inventor and thinker Heron of Alexandria wrote a piece on automata. Such machinery in Heron’s fictional accounts was usually activated with falling water, heat, or atmospheric pressure. According to another Greek myth, the inventor Daedalus, after building the Minotaur’s labyrinth,

made wings out of wax for himself and his son Icarus to escape Crete, but Icarus's wings melted when he flew too close to the sun. It is also said that Daedalus built a copper machine called Talos to protect Crete [2].

The Golem

In the late middle ages, alchemists experimented with an anthropomorphic figure, the Golem, who was made of clay and would be summoned when God was called upon. According to the mystical tradition, Golem was created by Rabbi Loew to protect Jews in seventeenth-century Prague from mistreatment. Like "Adam" in Hebrew (of earth), human Golem is made of clay. However, the legend highlights the uniqueness of the Holy one, who is the only one with true power to create life, because Golem was misshapen [4]. This idea comes through each time an anthropomorphized robot comes up, such as in Mary Shelley's nineteenth-century story, Frankenstein. The concept seems to fascinate the founders of Artificial Intelligence as well: John Von Neumann, Norbert Wiener, and Marvin Minsky claimed they are direct descendants of Rabbi Loew [4].

The Ars Magna

The Golem might have been a legend, but it had long-lasting psychic influence. During the Enlightenment, many had ideas of machines that could be used for mathematical operations as well as for the systematic production of ideas. It is interesting that the first effort for the systematic generation of logical statements was a long time before then.

The *Ars Magna* (Great Art) was invented in the thirteenth century by the pre-Reconquista Catalan theologian Raymond Lull (1232–1315) [5]. This tool could generate all combinations of a limited number of axiomatic principles, or concepts which were "true." These "true" axioms were desirable attributes in the Catholic tradition like eternity and goodness. The tool with two concentric circles contained the fourteen divine

attributes. The rotation of the circles could produce 196 twofold combinations of factors. By combining each element with every other, an extensive inventory of all true statements could be produced. The generation of combinations was not heuristic but syntactic. The tool gave all combinations, not just selected or constrained results. Other tools Lull made were for studying the seven deadly sins and other theological artifacts [5].

The Concept of Symbolic Languages

Hurufism

In the fourteenth century, a movement called Hurufism was established by Fazlullah Hurufi (d. 1394) [6]. As a unique religious movement, Hurufism (Makers of the Muslim World) influenced a wide area from India in the East to Anatolia and Balkans in the West. Fazlullah systematized the process of ascribing esoteric meanings to letters, which had already been done by his predecessors. After its founder's death, however, the Hurufism movement disintegrated, incorporating over time into other movements. Fazlullah's talent for interpreting dreams or the connections he drew between prayer sequences and someone's face or even between the number of joints a human has and their "degree" in the astrological world are some examples of his applications. Through these, Fazlullah made interpretations on topics no one had ever mentioned or considered and convinced his followers that he was a Messiah illuminating mysteries.

The Calculator

Inspired, in my opinion, by the Hurufism movement in the Islamic world, the first person to come up with essential symbolic calculation was the inventor of calculus and the inventor of the Stepped Reckoner calculator, Wilhelm Gottfried Leibniz (1646–1716) [7]. A most interesting theory Leibniz developed concerned monads, or atomic bits expressing the small aspects of given philosophical principles, introducing the

systematical use of symbols. Like Hobbes, who said “Reasoning is computation” one hundred years ago, Leibniz lacked the practical digital calculation and computation language to show that he had a good understanding of the concept. In the absence of any objective economic need for a computer, Leibniz and Hobbes’ suggestions for symbol processing remained more philosophy than computation. It was obvious that Artificial Intelligence required general purpose digital information processing and would have to wait until the twentieth century.

The Technological Myth Persists

AI and its early beginnings were accompanied by the technological myth of the creation of thinking machines through digital computing. Martin [8] mentions the image of the computer as an “awesome thinking machine.” When the digital revolution began around the 1950s and 1960s, many people saw the new computers as “intelligent brains, smarter than people, unlimited, fast, mysterious, and frightening.”

While the trend of “expert systems” in the 1980s resulted in profitable applications, new expectations about artificial intelligence also appeared. Around then, Japan launched a project called “Fifth Generation” to build intelligent machines over 10 years. The USA and UK had similar efforts, but none succeeded [9]. Developments in neural networks also led to a more connectionist approach to AI rather than symbol manipulation. The myth of Artificial Intelligence and its global influence could help us understand how such technological myths can impact the current digital media and its social presence [10].

In 1967, about 5 years after his work on “Fuzzy Sets” was published [11], Lofty A. Zadeh (born 1921), the founder of this mathematical theory saw the name of his department then at UC Berkeley change to incorporate “Computer Science” into its name: “Electrical Engineering and Computer Science”. He wrote: “What we still lack, and lack rather acutely, are methods for dealing with systems which are too complex or too ill-defined to admit of precise

analysis. Such systems pervade life sciences, social sciences, philosophy, economics, psychology and many other “soft” fields [11]. In the first years after his discovery, Zadeh’s theory was intended to be applied to other fields like humanities and the social sciences. In a 1994 interview in the *Azerbaijan International*, Zadeh was asked, “How did you think Fuzzy Logic would be used at first?” his retrospective answer was, “In many, many fields. I expected people in the social sciences economics, psychology, philosophy, linguistics, politics, sociology, religion and numerous other areas to pick up on it. It’s been somewhat of a mystery to me, why even to this day, so few social scientists have discovered how useful it could be.” [11]. Zadeh underlined that the fundamental principles of “thinking machines” had been developed by mathematicians.

AI, a research program launched in 1959, has now become widespread, and more or less successful at times, in the fields of science and technology, but after half a century, it would not be wrong to say that AI has not met the world’s expectations. Take chess as an example. Chess has always been considered a game of intellect, and many computing pioneers have said that machine that could play it well would display true artificial intelligence. Machine intelligence is usually challenged with the Turing Test but chess can also be a good test. In fact, the ability to play chess was even used as a valid question to ask during a Turing Test in Turing’s original paper. AI researchers have created programs that even beat the world’s best chess players. However, this does not mean that the best machines understand the concepts of chess, only that they use plays well.

AI is on its Way

Ironically, the history of information processing is quite mysterious. While presently the Internet enables us to reach more information than we could ever need, computers were actually developed within a policy framework that limited information flow and only a limited number of people knew about them at the time.

Solutions are sought for the early diagnosis and treatment of the diseases, especially in regions

Table 1 The development of AI as we know it today

Period	Stage	Highlights
1940–1950	Gestation	McCulloch & Pitts: Boolean circuit to model of brain Turing's computing machinery and intelligence
1950–1970	Early enthusiasm, great expectations	Early AI programs, Samuel's checkers program Birth of AI @ Dartmouth meeting 1956
1970–1990	Knowledge-based AI	Expert systems, AI becomes an industry
1991	Today and future	AI becomes an essential partner in all areas of our lives

where access to physicians is difficult. Verifying the reliability of the system can become a greater challenge, as it requires verifying what the system is trying to do rather than verifying the safe behavior of the system in all areas in which it operates. Table 1 shows the stages past, present, and predicted regarding AI, based on its features, starting from its inception [12].

High capacity electronic calculators were actually born as military inventions to decrypt enemy messages or to calculate missile routes and were thus highly protected to ensure no details got leaked. It took several decades for the British government to completely de-classify secret documents involving the digital computing machine used during the war, developed at Bletchley Park military base [13] (as will be explained in greater detail below). Electronic information is able to wander the earth over the Internet as it pleases but it is also surrounded by a web of secrecy.

In the mid-twentieth century, as science became more and more militarized, two conflicting ideologies appeared. Scientists believed in the free exchange of information (at least in principle), which could hasten progress. On the other hand, governmental intelligence agencies broke down activities into small cells, each with limited information. These differences in opinion, each unconditionally adhered to by its proponents, turned into severe conflicts when military commanders began heading war projects involving scientific research such as the atom bomb and electronic computers. Instead of sharing their results at international conferences, scientists chose to comply with the limitations introduced for the sake of national security and kept their work a secret.

This culture of secrecy in computer science continued with the Cold War, during which secret research was done on defense systems. In an effort

to establish electronic dominance over the Russians, the American government spent a great deal of money not only on its armed forces but also on private companies and universities producing computers. Military, academic, and commercial interests were enmeshed. The computer called Harvard Mark III was designed for the US Navy and produced in a university laboratory under the sponsorship of IBM.

Military efforts clearly had an important role in the advent and development of computer and related technologies. War is truly horrific, but it must be acknowledged that continued efforts to build a nation and a powerful military force can result in important technological innovations, especially general-purpose technologies for new ways of processing material or information [14]. These technologies can result in changing entire sectors. Some examples concern energy creation, information processing such as steel replacing iron, autoes replacing horse drawn carriages, and the computer replacing other older tools. These technologies will be accompanied by new forms of communication, activities, and many ancillary industries. Therefore, the roots of AI, the digital computer, portable music players, smart phones, and many software applications can be traced back to World War II. The Cold War was also paradoxically instrumental in shaping the Computer Revolution [15].

Conceptual Revolutions

Until 1930–1950, intelligence was viewed as having a more tangible quality rather than having qualitative aspects to it, thus necessitating a drastic change. The new focus meant that the new goals needed to be clarified and planned. Like Paul Thagard [16] mentions in “Conceptual

Revolutions,” a total paradigm change was to happen with the adoption of Artificial Intelligence as a fundamental science and engineering problem, causing an overall shift in focus.

At that time, it seemed like a win-win situation: The private companies were getting funding from the government and making it through the hard times, taking advantage of a guaranteed market. Meanwhile, military experts had immediate access to the newest products. On the other hand, scientists without access to government funding had great difficulties accessing the benefits of information processing. And those who did accept funding support could not be ethical and transparent, having to comply with their employers’ condition of nondisclosure [17].

During WW2, the military inventors for Britain, Germany, and the USA worked in the computer science field secretly and separately. In 1942, physicist John Mauchly was working on the all-electronic calculating machine as the US Army needed a device that would calculate complex wartime ballistics tables. Within a year, Mauchly began building the Electronic Numerical Integrator And Computer (ENIAC) and took 2 years [17]. Civilians only got wind of these developments a year later, in 1946 when the US military announced it. This machine was built in a university laboratory, but its production took place under military oversight. Until a lightning strike in 1955, the ENIAC might have carried out more calculations than all of humanity had until then [17].

The media presented the ENIAC as the world’s first electronic computer. It took up a whole room, but it would not compare to a present-day laptop in terms of speed and power. The newspapers still announced: “The devices in this machine” worked “faster than the neurons in the human brain.” [18]. It often broke down and had to be fixed by humans. It emitted heat, and moths and flies would get in and ruin the connections. The term “debug” can be traced back to these first pests. A more serious issue was its limited capabilities. The ENIAC was actually designed to calculate ballistic tables, but anything else like the weather or the movement of shock waves required elaborate rewiring, which would take days and was

usually done by women. The ENIAC was more of a giant calculator than a precursor to a computer and it made no sense to ask it to calculate something else without physically rebuilding it [18].

World, Meet the Personal Computer

Known as the world’s first digital electronic computer, the Atanasoff-Berry Computer (ABC) was built between 1939 and 1941 by electrical engineer, mathematician, and physicist John Vincent Atanasoff and his assistant, physicist Clifford Edward Berry. Atanasoff established the four main ideas on which the digital electronic computer is based [19]:

- 1) The use of electricity and electronic technologies to ensure speed in computer environments,
- 2) using the binary system to simplify the computing process,
- 3) using renewable memory to decrease production cost, and
- 4) computing using direct logical processes rather than counting to increase accuracy.

Around that time, known only by a small group of British scientists at Bletchley Park, another military project was underway to build functional machines called the Colossus that would aid in decrypting German intelligence. Secrecy was crucial, since the entire project depended on the German’s remaining unaware that the British were decoding diplomatic messages and acting according to these even if the Germans changed the encryption daily. Speed was also crucial. If a German submarine or air strike was to be preempted, the order had to be decrypted before the encryption changed. For security purposes, the staff at Bletchley worked in small groups and was only informed of their immediate task. Thousands of men and women kept true to their vows of secrecy, never revealing that the first digital computer had been built not by the Americans but by the British [20].

What set the Colossus apart from the ENIAC was that it could make choices. By the end of the war, 10 electronic Colossus machines worked on messages intercepted, compared them to countless message samples, found patterns and similarities and suggested a way of encrypting that day’s

secret code. It had been this ability – that could choose – that ensured the success of this mission. Instead of scanning all probabilities without thinking, it followed the previous commands put into it or paused for a moment to ask questions of the human operator and quickly eliminated numerous dead ends. The entire base functioned as a giant information processing machine, receiving unintelligible messages, and produced comprehensible details about the Germans' plans. The human, mechanical, and electronic components in the machine followed certain instructions and interacted. Unbeknownst to the many people there, one of the world's most prominent mathematicians in the field of decision making, Alan Turing, had also worked at Bletchley Park. Today, Turing is famous for founding the modern information society on which the global communication of power and money control depend [21].

The Big Question: Can Computers Think?

With Turing and his colleagues keeping their wartime activities a secret, the *first programmable computers* were invented without anyone knowing. Regardless, Turing's influence was immense, since he had not only thought up the technology that made computers work, but also the implications of this technology. After the war, military and commercial institutions focused on producing bigger, faster, stronger, and most importantly, easily programmable computers that could shift between tasks. Turing posed some fundamental questions about *machine intelligence* and brought the *human mind and electrical circuits* closer than ever. Since his closest friend had died at a young age, he had a skeptical view of the traditional Christian concept of the soul, a view that was reflected in machine philosophy. Like biological determinists who sought the secrets to life in complex molecules, *Turing thought computers might be able to think despite their make-up of electrical circuits*. Though he acknowledged thinking was a hard act to define, he was convinced that humans and computers could think.

Agreeing with Turing's view meant having a new conception of both *humans and machines*. Were computers modeled and designed after the human brain or was it the opposite? The scientists who, at the beginning, were excited to proclaim that these circuits resembled super-fast neurons, quickly claimed the nervous systems of living things functioned just like electronic systems. According to their conception of the human soul, humans made decisions as a result of signals moving like electronic buttons that chose one of two ways, zigzagging through branches. Typists were a popular example. It was said that when the boss was dictating, the secretary's ears would catch the sound waves and the body/brain would decrypt these and turn them into simple electronic signals, which in turn activated the fingers (computer enthusiasts would take it further to add that the secretary could enter the memory storage part to correct grammar mistakes) [21].

In Turing's projects for the future, real and mental experiences were intertwined. In the 1930s, at a time when it was not possible to physically build electronic computers, Turing had invented a machine that took instructions from a long strip of paper filled with signs and spaces and claimed that his *machine behaved just as a human would*.

Earliest Steps towards Computerized Medicine

In the eighteenth century, when **Giovanni Battista Morgagni** (1682–1771) published the magnificent “*De sedibus et causis morborum per anatomen indagatis*,” or “Anatomy Rules,” he could not have known that he had just laid the groundwork for *computerized medical diagnosis*. His volumes and their indices resulted in the first true encyclopedia accurately correlating pathological and clinical information. This would provide immensely useful in assisting pathologists and other physicians as they tried to diagnose diseases [22].

The indices of Morgagni's “*De Sedibus*” were perhaps the first step toward the automation of medical information. Then, a perforated-card

mechanism was used in the weaving machine of Morgagni's contemporary, Jacquard, building upon the work by Bochon (1725), Falcon (1728), and Vaucanson (1745). Punched cards would be used as a method to store and sort data for later analysis. About a century later, in the early 1850s, Dr. Joseph Henry, the first Smithsonian Institution Secretary, foresaw the need for new methods of classification and correlation of information as he observed that existing methods could not handle the vast scientific data compiled at that time [23]. Jumping forward to the 1930s and 1940s, the more rapid accumulation of data accumulated in medical research pushed an increasing number of researchers to use punched cards as mechanical aids to classify and correlate data in their own fields. In 1952 punched cards were used to automatically correlate data in the differential diagnosis of hematologic diseases, and in 1961 a computer was introduced into medicine for that purpose [22]. The many great advances that followed in the application of computer technology, in so many areas of medical research and practice, have led to revolutionary improvements in bibliographic, laboratory, radiologic, and other branches of medicine and have fulfilled expectations of those active in the field, who could barely have been imagined several decades ago [22].

In short, the Morgagnian correlation exemplified in the indices of "De sedibus," and the later development of automated methods of correlation of medical information that arose in the 1950s and 1960s set the foundations for using computers in medical diagnosis [22]. The first large-scale application of automated techniques in medical practice was carried out by Collen [24] in his multiphasic screening program. This was broadened by the application of electronic data processing systems to many medical specialties.

Collaboration Between Medical Scientists and AI Researchers

The first efforts towards collaboration in the field of life sciences took place in the USA between computer scientists and medical scientists and led

to computing resources such as the *Advanced Research Projects Agency Network (ARPANET)* and the *National Science Foundation Network (NSFNET)*, allowing researchers to address medical problems using AI methods. In 1985, as awareness that computers could be useful in the clinic increased, the first *Artificial Intelligence in Medicine (AIME) Conference* took place, bringing together computer scientists, doctors, and biologists among others. In 1989, a journal named *Artificial Intelligence in Medicine* was founded to advance research within the field [25]. Many existing scientific gatherings often incorporated presentations on different aspects of machine learning. More conferences were also held on machine learning in health care such as "Artificial Intelligence in Medicine (AIMed)," which holds separate conferences in Europe, North America, and Asia, the "Human Intelligence and Artificial Intelligence in Medicine" symposium, and "Machine Learning in Healthcare" meeting at Stanford University, the "Deep Learning in Healthcare Summit" in Boston, "Machine Learning, Big Data and AI in Healthcare" conference in Washington, and the "Predictive Analytics World for Healthcare" in Las Vegas.

Examples of the Use of AI in Medicine

Mirroring the developments in other industries, the use of artificial intelligence in the field of medicine is on a steady increase. Major companies in various medical sectors, notably the pharmaceutical and imaging sectors, have made large investments in this field. Artificial intelligence software research has also attracted a great deal of academic interest. While there are many publications on artificial intelligence applications in different medical fields demonstrating the breadth of the uses of these techniques, those approved are not as many.

One important example of artificial intelligence applications used in medicine is DeepVariant, the artificial intelligence framework announced in 2016 by Google. It is able to identify single nucleotide polymorphisms, the most common genetic variation, with 99.9587%

accuracy, for which it was awarded by the FDA [26]. Then, there are applications for assessing suspected malignant melanoma based on skin lesion photographs and detecting lymph node metastasis of breast cancer by analyzing pathology slides [27]. Another example is a radiology algorithm developed at Stanford University that was able to diagnose pneumonia more accurately than radiologists [28]. Finally, in the field of ophthalmology, many artificial intelligence studies have been successful in numerous digital techniques like color fundus photography, optical coherence tomography (OCT), and computerized visual field (VF) testing and the huge databases they provide. As populations live to an older age, there have been an increase in eye diseases that cause preventable vision loss. The artificial intelligence applications underway particularly for diabetic retinopathy (DR), age-related macular degeneration (AMD), glaucoma, and retinopathy of prematurity (ROP), as the leading causes of vision loss will be of immense use [12, 29]. More and more examples come up every day in various medical fields.

The progress made in artificial intelligence studies makes it clear that potential future applications are promising. AI holds great potential for identifying patients with preventable diseases and referring them to a physician, especially in developing countries where health-care access is challenging [12].

Artificial and Biological Life

In terms of aspects of humanity, artificial and biological life can be compared as much for rational thinking as for feelings and emotions. In fact, there have also been attempts to anthropomorphize computers and robots in terms of how computers learn the way children do, theorizing that computer might also learn through experiences and trial-and-error [30].

One end of the AI myth spectrum is occupied by the Apocalypse, but there are more moderate, utopian theories as well. As personal computers spread and the Internet became a daily presence in our lives, many new technological myths

emerged, involving new forms of the myth of AI, these days usually around collective intelligence and networks. The new themes explored by authors within the myth in “*networking AI*” feature utopian visions based in previous telecom advances, with the Internet seen as the final stage of human interconnectedness, leading to interactions between humans and machines collaborating to increase collective intelligence to new levels. The new global brain could offer humans a new level of consciousness [31]. Pierre Levy, the media theorist, says traditional AI is limited and suggests a new concept he calls a collective “*Hypercortex*.”

The patterns that characterize the AI myth dating back to the 1940s to the 1970s are visible in contemporary versions of the AI myth, too. Technology is both humanized and associated with superhuman or even supernatural powers [32]. We have yet to see which ones will come true within our lifetimes.

Conclusion

C. W. Mills [33] says that each era has its own mindset. This mindset, while not the only one, is the most important sign of that era, as it contains our strongest beliefs and perspectives of the time. It is also a summary of our most accurate senses and vulnerabilities. It is the organization of the senses around this sign as the manifestation of a social utopia. Simply put, it is a visualization and realization of society and social relations. Sociological thought, or the ability to create a social dream, can organize within itself, the farthest (such as machines) with the nearest (such as the human mind) through other and new senses to construct a self-sufficient, harmonious, and inclusive vision.

The increasingly extensive and complex computer programs that control human life in a myriad of ways are becoming harder to understand. Applications that have the ability to transform themselves are slowly appearing and will become more widespread and even harder to understand. The unpredictability of the outcomes of such programs is an important problem.

Developing machines that we cannot have complete control over and handing over the most sensitive aspects of our lives to these machines are some of the greatest risks humanity has ever taken [34].

We are witnessing research into intelligence extend into fields that require expertise like “*computing, decision making, memory, learning, creating, planning*” towards the production of machines that can outdo the human capacity to think. Given that machines can transmit all that they have learned to the next generation, machines could have an evolutionary advantage against humans [34]. As humans do this transmission using external mediators, only part of the information and experience transmitted can be internalized by the next generations.

Zadeh [35] did not believe that “*thinking machines*” thought the way humans do, so one aspect of his work involved “*making computers think like people*” in the mid-1980s. He suggested that the machine’s ability “to compute with numbers” was supplemented by an ability like human thinking. In 1990, he started developing a scientific concept he called “*soft computing - and, in particular, fuzzy logic - to mimic the ability of the human mind to effectively employ modes of reasoning that are approximate rather than exact.*”

For over 30 years, AI has been expected to have a revolutionizing influence on medicine. There have indeed been important technological developments recently that could have enormous effects, including exponential increases in computing power, big data processing technologies, access to large clinical data sets using electronic health records, and machine learning [36]. These could allow doctors to make more accurate diagnoses due to improving algorithms and also individualize patient treatment [37].

Science and technology are like the proverbial shining sun, as a constant source of light. However, many tests are still needed to ensure the safety of artificial intelligence applications, like any technological advances directly affecting healthcare in particular [38]. The way to ensure the safety of humankind against any risks in developing AI is to establish well-thought-out

restrictions and regulations and to work within an ethical framework. The development of new ethics, laws, and a philosophy compatible with our new life is crucial.

It is important for physicians to become more familiar with the basic concepts and metrics of machine learning and their potential applications and follow the increasing integration of artificial intelligence and machine learning into modern medicine. We can already see that artificially intelligent robots are going to evolve in unique and peculiar ways, which are currently hard to predict. One thing we want to avoid is rendering the robotic difference into a stigma, in terms of how far this difference can be from the human norm. The difference may become an evolutionary trait of existence. The integration of the potential of humans and robots in the future could lead to an *onto-epistemological approach* and ultimately develop into an *existential quest* for the original interspecies [39].

A more recently popularized term sometimes equated with artificial intelligence is “*deep learning*,” which is actually a tool included in the many offered by various statistical and probabilistic algorithms. With “*learning*” now possible, the smallest data put into algorithms could result in unforeseeable consequences, making developers and designers responsible for preventing unexpected and unforeseeable actions by AI systems. Overseeing resources and projects should be diverse boards comprising academics and professionals to ensure safety. The need and rights of humans and patients such as autonomy should always be taken into consideration since the probable outcomes and the risks of certain medical interventions, especially neurological ones, involve uncertainties [38]. Effective safety measures can only be possible through continuous ethical evaluations and monitoring technological development.

Let us hope, as Aristotle says in first book of *Nicomachean Ethics*, [40] that “*Every art and every kind of inquiry, and like-wise every act and purpose, seems to aim at some good: and so it has been well said that the good is that at which everything aims.*”

References

1. Skinner RE. Building the second mind: 1956 and the origins of artificial intelligence computing. Berkeley: California University Publish; 2012. p. 14–8.
2. Plato. Dialogues of Plato. New York: Simon & Schuster; 2010.
3. Polya G. How to solve it. Princeton: Princeton University Press; 2004 (first print. 1945). 3,32,51 pp.
4. Nicole D. Interview: Doron & Yoav Paz for the Golem – nightmarish conjurings [Internet]. nightmarish-conjurings.com; 2018 [cited 2020 Dec 10]. <http://www.nightmarishconjurings.com/2018/11/05/interview-doron-yoav-paz-for-the-golem/>
5. Aspray W, editor. Computing before computers. 5th ed. Iowa State University Press; 1995. p. 104–6.
6. Ballı HH. Hurufilik nedir? (What is Hurufism?). [Internet]. e-makâlat Mezhep Araştırmaları, 4(2);2011 [cited 2020 Dec 10]. 31–48. http://isamveri.org/pdfdr/G00004/2011_2/BALLI.pdf
7. Computer History Museum. The Leibniz step reckoner and curta calculators. [Internet] [Cited 2020 Dec 10]. <https://www.computerhistory.org/revolution/calculators/1/49>
8. Martin CD. The myth of the awesome thinking machine. Commun ACM. 1993;36:120–33.
9. Russell SJ, Norvig P, Canny JF. Artificial intelligence: a modern approach. Saddle River: Pearson Education; 2010. p. 24–5.
10. Natale S, Ballatore A. Imagining the thinking machine: technological myths and the rise of artificial intelligence. Convergence. 2017;1–16. <https://doi.org/10.1177/1354856517715164>.
11. Zadeh LA. Fuzzy sets, information and control. 8th ed; 1965. p. 29–37, 338–353.
12. Keskinbora KH, Güven F. Artificial intelligence and ophthalmology. Turk J Ophthalmol. 2020;50:37–43. <https://doi.org/10.4274/tjo.galenos.2020.78989>.
13. Marsh A. The hidden figures behind Bletchley Park's code-breaking colossus [Internet]. 2019 [cited 2020 Dec 10]. <https://spectrum.ieee.org/tech-history/dawn-of-electronics/the-hidden-figures-behind-bletchley-parks-codebreaking-colossus>
14. Adams IV NA. Creating clones, kids & chimera: liberal democratic compromise at the crossroads. 17 Notre Dame J.L. Ethics & Pub. Pol'y71; 2003. <http://scholarship.law.nd.edu/ndjlepp/vol17/iss1/4>
15. Dolsma W, van den Ende J. Technology push, demand pull and the shaping of technological paradigms – patterns in the development of computing technology. J Evol Econ. 2004;15(1):83–99. <https://doi.org/10.1007/s00191-004-0220-1>.
16. Thagard PR. Conceptual revolutions. Princeton: Princeton University Publication; 1992. p. xvi.
17. Campbell R, Martin CR, Fabos B. Media & culture 2016 update: mass communication in a digital age. 10th ed. Boston: McMillan; 2016.
18. Computer History Museum. ENIAC. [Cited 2020 Dec 10]. <https://www.computerhistory.org/revolution/birth-of-the-computer/4/78>
19. Atanasoff JV, Brandt AE. Application of punched card equipment to analysis of complex spectra. J Opt Soc Am. 1936;26:83.
20. Turing AM. Computing machinery and intelligence. Mind. 1950;LIX:433–60.
21. Clement A. Looking for the designers: transforming the ‘invisible’ infrastructure of computerised office work. AI & Soc. 1993;7:323–44.
22. Lipkin M. Historical background on the origin of computer Medicine. Proc Annu Symp Comput Appl Med Care. 1984;987–990. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2578563/>
23. Henry J. Annual report of secretary. Washington, DC: Smithsonian Institute; 1851.
24. Collen MF. Periodic health examinations using an automated multitest laboratory. JAMA. 1966;195:830–3.
25. Patel VL, Shortliffe EH, Stefanelli M, et al. The coming of age of artificial intelligence in medicine. Artif Intell Med. 2009;46:5–17.
26. Google Genomics. Precision FDA truth challenge [Internet]. [Cited 2020 Dec 10]. https://googlegenomics.readthedocs.io/en/staging-2/use_cases/discover_public_data/precision_fda.html
27. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115–8.
28. Kubota T. Algorithm better at diagnosing pneumonia than radiologists [Internet]. Stanford Medicine News Center. [Cited 2020 Dec 10]. <https://med.stanford.edu/news/all-news/2017/11/algorithm-can-diagnose-pneumonia-better-than-radiologists.html>
29. Flaxman SR, Bourne RRA, Resnikoff S, Ackland P, Braithwaite T, Cicinelli MV, et al. Vision loss expert group of the global burden of disease study. Global causes of blindness and distance vision impairment 1990–2020: a systematic review and meta-analysis. Lancet Glob Health. 2017;5: e1221–34.
30. Robertson M. What computers can learn from children. New Scientist. 1975;68:210–1.
31. Das HF. Globale Gehirn als neue Utopia. In: Maresch R, Rötzer F, editors. Renaissance der Utopie: Zukunftsfiguren des 21. Jahrhunderts. Suhrkamp: Frankfurt am Main; 1975. p. 92–112.
32. El Kalouby R, Robinson P. Mind reading machines: automated inference of cognitive mentalstates from video. In: 2004 IEEE international conference on systems, man and cybernetics. IEEE; 2004. p. 682–8.
33. Mills CW. Toplumbilimsel düşün (trans: Oskay Ü). Ankara: Kültür Bakanlığı; 1979. p. 24.
34. Georges TM. Digital soul: intelligent machines and human values. Boulder: Westview; 2003. p. 6–7, 11.
35. Zadeh LA. Making computers think like people. IEEE Spectrum. 1984;8:26–32.
36. Sadegh-Zadeh K. In dubio pro aegro. Artif Intell Med. 1990;2:1–3.
37. Weiss J, Kuusisto F, Boyd K, Liu J, Page D. Machine learning for treatment assignment: improving

- individualized risk attribution. AMIA Annu Symp Proc. 2015;2015:1306–15.
38. Keskinbora KH. Medical ethics considerations on artificial intelligence. J Clin Neurosci. 2019;64:277–82. <https://doi.org/10.1016/j.jocn.2019.03.001>.
39. Ferrando F. Is the post-human a post-woman? Cyborgs, robots, artificial intelligence and the futures of gender: a case study. Eur J Futur Res. 2014;2:43–60.
40. Aristoteles. Nikomakhos'a etik. İstanbul: Bilgesu Yayıncılık. 2000; 17p.



AIM and Patient Safety

14

M. Abdulhadi Alagha, Anastasia Young-Gough,
Mataroria Lyndon, Xaviour Walker, Justin Cobb,
Leo Anthony Celi, and Debra L. Waters

Contents

Introduction	216
Trends in Quality and Safety Research	217
Approaches to Patient Safety and Preventing Errors	217
Intelligent Systems: Machine Learning and Natural Language Processing 218	
Prevention of Adverse Events	219
Diagnostics	219
Medication Errors and Polypharmacy	220
Treatment Outcomes and Quality	221
Patient Safety Databases	222
Future of AI in Safety and Quality	222
References	223

M. A. Alagha

MSk Lab, Department of Surgery and Cancer, Imperial College London, London, UK

Institute of Global Health Innovation, Department of Surgery and Cancer, Imperial College London, London, UK
e-mail: h.alagha@imperial.ac.uk

A. Young-Gough
Preventive and Social Medicine, Dunedin School of Medicine, Dunedin, New Zealand

M. Lyndon
Centre for Medical and Health Science Education, University of Auckland, Auckland, New Zealand
e-mail: mataroria.lyndon@auckland.ac.nz

X. Walker · D. L. Waters
Department of Medicine, University of Otago, Dunedin, New Zealand
e-mail: xaviour.walker@otago.ac.nz;
debra.waters@otago.ac.nz

J. Cobb

MSk Lab, Department of Surgery and Cancer, Imperial College London, London, UK
e-mail: j.cobb@imperial.ac.uk

L. A. Celi (✉)

Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA

Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, MA, USA

Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA
e-mail: lceli@mit.edu

Abstract

Patient safety has constituted a huge public health concern for a long period of time. The focus of safety in the healthcare context is around reducing preventable harms, such as medical errors and treatment-related injuries. COVID-19 pandemic, if anything, has act as a wake-up call for health experts to address latent safety problems. Advancements in the field of artificial intelligence have highlighted the use of intelligent systems as a proven means of improving patient safety and enhancing quality of care.

This chapter explores trends in quality and safety research, the use of machine learning and natural language processing in the context of improving patient safety and outcomes, the use of patient safety databases as a source of data for machine learning, and the future of artificial intelligence in quality and safety.

Keywords

Patient safety · Quality and safety · Medical errors · Treatment injury · Adverse events · Polypharmacy · Natural language processing · Machine learning

Introduction

Patient safety has represented a serious global public health concern for centuries. As the worldwide health systems come under increasing strain and burdens surge, medical errors and treatment injuries will make up for a growingly large proportion of the world's total morbidity and mortality [1]. The complexity of healthcare systems involves a mismatch between "patient guidelines" and "patient needs" and arguably increases the probability of errors and suboptimal patient care, rendering safety improvement a major challenge facing modern healthcare systems. Patient safety is defined as the absence of preventable harm and medical errors to a patient and the prevention of potential harm risk associated with the healthcare

process. Central to this is the prevention of medical errors and treatment injuries [1].

Medical errors are reported as the third leading cause of death in the United States and are process faults that have the potential to cause adverse patient outcomes and can occur at any stage on the continuum of care [2]. Treatment injury (TI), on the other hand, is the personal injury caused by an accredited health professional as a result of delivering treatment [3]. Most medical errors are a result of system- or process-based failures, and not an individual's mistake [2]. Shojania et al. [4] recommended 11 practices as "clear opportunities for safety improvement"; however, patient-related factors are also surrogate markers of patient safety and include accurate detection of treatment injuries, adequate and effective treatment and recovery, and high function in the mid to long term [5].

Over the past decade, rapid advancements in artificial intelligence (AI) techniques have sparked a revolution in the medical sector as a possible means of delivering optimal patient care, and there has been an abundance of evidence to support the effectiveness of such technology in improving quality and safety measures across a variety of use cases. Leveraging AI with medicine is not only aimed at transforming the roles of healthcare providers but is also thought to provide a patient safety platform to eliminate medical errors and adverse events and enhance quality of care and safety outcomes [6]. The AI Clinician system was demonstrated to provide personalized and clinically interpretable sepsis treatment decisions to assist human clinicians [7]. An AI algorithm was similarly robust in detecting breast cancer using mammogram images [8]. Several studies showed the role of AI in improving workflow and reducing medical errors in different medical fields [9].

However, given the limited availability of data repositories in the safety and quality domain, we sought to present the potential use of AI techniques to enhance patient outcomes. The following sections draw from recent advancements as well as existing gaps in quality and safety research in order to highlight potential AI strategies in this field.

Trends in Quality and Safety Research

The trend toward electronic health records (EHR) and electronic-based compulsory audit software has given rise to large structured and unstructured datasets of patient information. The electronic health record has been regarded both as an advancement in quality and safety and as a resource for quality and safety research [10].

There has also been a move toward the privatization of quality and safety through the creation of large-scale quality and safety improvement entities such as the American College of Surgeons National Surgical Quality Improvement Program (ACSNSQIP) and the use of commercial audit software to meet the audit requirements of professional regulatory bodies [11]. This software gives the clinician specific data points to fill in for each patient audited in order to generate structured data points that are more interpretable and comparable to other centers using the same system. As quality and safety is now a business venture in medicine, it is further incentivized to improve.

In the field of primary care, there has been a move to greater accessibility of EHR which has been proposed as a way of improving patient satisfaction alongside other tenets of quality of care (i.e., efficacy, efficiency, equity, and safety). A recent systematic review and meta-analysis found that EHR sharing improved HbA1c levels in type 2 diabetics [12]. The encouraged sharing of EHR could also contribute to the accuracy of these records by allowing patients to see and correct the recorded information [13]. Improvements in primary care have often been proposed as a solution to the rises in emergency department (ED) attendances, and therefore, avoidable ED attendances have been used as a marker of primary care quality. As a marker, this has been shown to be sensitive to care quality, but the magnitudes of association are small [14].

There has been a cultural change in medicine shifting the focus of quality and safety practices around adverse event reporting from the individual practitioner who errs to the failing of systems and societal impacts [15]. The All New Zealand Acute Coronary Syndrome Quality Improvement (ANZACS-QI) program, is a New Zealand

Ministry of Health (MOH) endorsed nationwide audit, with completion rates built into the annual MOH reporting requirements, and as such, has a high level of capture of Acute Coronary Syndrome (ACS) data [16]. Linking audit requirements into a government funding body ensures the audit has a high level of capture, and also allows the dataset to be linked to the routinely collected New Zealand National Hospitalisation Dataset. Audit software such as this both collects and compares data for future quality and safety improvement, and also provides a platform for additional research [17]. As the ANZACS-QI program is standardized and compulsory for all New Zealand hospitals with patients that meet the criteria, additional areas of interest can be added to the system to look at further areas of research. For instance, the national study on oxygen therapy in ACS, a data field for oxygen was added and made compulsory on randomized days to gain data on usage in the setting of ACS [18].

Similar to the medical specialties, surgery is also subject to compulsory auditing as part of governmental influence and specialty college requirements [19]. The current model for surgical quality assessment and improvement is largely reliant on the characteristics of larger hospital centers, including the presence of other providers for peer review, higher procedure volumes in the center, and resources for data collection systems [20]. For those centers that can afford it, quality improvement programs such as the ACSNQIP and the Otago Clinical Audit provide a method of comparing clearly defined data against the expected standard and targeting quality improvement efforts. These programs also can work synergistically with other quality improvement efforts such as the traditional morbidity and mortality meetings which have been shown to underestimate postoperative complications [21].

Approaches to Patient Safety and Preventing Errors

In 2000, Reason conceptualized error into two broad models; the person-centered approach and the system-based approach gave a rise to

different philosophies of error management. Patient-centered approach accepts that unsafe acts are egregious and that individuals make error drawing on the unwarranted variability in human behaviors. This approach emphasizes the importance of disciplinary measures and training to standardize human acts. The system-based approach focuses on the conditions under which individuals work and aims to build defense mechanisms and layers to minimize errors and mitigate their consequences. This is better explained in Reason's Swiss cheese model of system accidents wherein slices of Swiss cheese have many holes that appear due to both active failures and latent conditions. The theorem argues that these holes are constantly opening, closing, and shifting position according to conditions and it is only when these holes line up to form a trajectory that system safety defenses are breached [22]. This model acknowledges that the pursuit of safety is about changing organizational culture to learn and adapt from occasional setbacks to develop a robust and feasible system that functions in the adversity of human and operational hazards.

Traditional approaches to safety management focus on error reduction using a "find and fix" identification lens and was later translated into the Safety I approach [23]. The aim of Safety I is to minimize error and eradicate harm while accepting the inevitable fallible reality of humans. Unlike Safety I which focuses on minimizing adverse outcomes, Safety II approach recognizes that safe care is provided in 90% of the cases and thus aims to encapsulate and maximize the number of acceptable outcomes. Safety II approach perceives humans as flexible agents of complex change and examines positive outcomes to invest in spreading good practice and knowledge building [24].

The following section highlights the role of AI to provide an alternative perspective of learning that harmonizes with traditional safety approaches and develops resilient healthcare systems.

Intelligent Systems: Machine Learning and Natural Language Processing

AI in medicine is the ability of a computer to make intelligent clinical decisions through analyzing extensive health data; identify trends, risks, and hidden knowledge; as well as enhance the interpretation of clinical outcomes. Among different types of AI, natural language processing (NLP) and machine learning (ML) specifically have promise in the safety and quality domains [25].

Machine learning enables computer systems to utilize algorithms and statistical models without explicit programming in making predictions and identifying latent knowledge from both labelled (supervised learning) and unlabelled data (unsupervised learning). Unlike ML-structured data, natural language processing enables computers to understand and translate human language (unstructured data) to machine-readable structured data which can be then analyzed using ML techniques.

The computerization of EHR to facilitate prevention, diagnosis, patient management, and treatment is still limited because they require access to reliable, coded, and retrievable clinical data. While many patient safety initiatives have been targeting structured data using ML techniques, textual reports of patient encounters include a wealth of unfiltered clinical information. A major challenge of implementing NLP in medicine lies in the diverse writing styles of healthcare professionals as well as the different contextual meanings of words. For instance, if a physician wants to identify patients with sepsis using the notes, they will retrieve too many patient reports containing expressions like query sepsis, sepsis excluded, rule out sepsis, and sepsis cannot be excluded. Although NLP systems have the ability to overcome these challenges to extract and structure relevant text-based clinical information, they require substantial combined medical and engineering knowledge and expertise to develop.

Prevention of Adverse Events

The definition of an adverse event varies widely in the literature. The World Health Organization (WHO) defines adverse events as “An injury related to medical management, in contrast to complications of disease” [26]. Worldwide, adverse events are estimated to affect 10% of hospitalized patients, in which 5–21% of these adverse events result in death and 50% are thought to be preventable with an estimate annual cost in the United States of \$17.1 billion [27]. Previous literature showed that the three most common types of adverse events are related to surgical complications, medications, and healthcare-associated infections [28]. In terms of cost and incidence, pressure ulcers followed by postoperative infections are the most common and costly, followed by central venous catheter infections and infections following infusions [28]. Healthcare has adapted from the aviation industry strategies such as the use of checklists for procedures, care bundles for central venous line insertion, and medication reconciliation.

Although there are national adverse event registries, such as the US Food and Drug Administration (FDA) Adverse Event Report System (FAERS), most hospital systems rely on retrospective chart reviews. Two common methods for detecting adverse events through retrospective chart review include the “Harvard Medical Practice Study” method and the “Global Trigger Tool” method, which have an overall positive predictive rate of 40.3% and 30.4%, respectively [29]. Both of these methods are multistage reviews that require a team of investigators and are time intensive. With the increased availability of EHR and development of NLP techniques, hospital systems and researchers are starting to validate models for accurately detecting high-impact adverse events such as hospital-acquired and ventilator-associated pneumonia and central line-associated bloodstream infections (CLABSIs). Using data mining methods, CLABSIs have been shown to be predictable (and therefore preventable),

significantly improving clinical surveillance programs, reducing hospital length of stay, and advancing patient safety. In improving postoperative infection rates, models using machine learning and AI have been used in identifying spinal fusion infection. With regard to preventing pressure ulcers, deep learning approaches have been developed using wearable computing sensors to alert care providers when a patient is in position for too long a period. Machine learning has also been used to identify which patients would benefit from limited speciality beds to reduce pressure injuries in the intensive care units.

Pharmacovigilance offers a significant opportunity for AI to detect adverse events, with feasibility studies showing effectiveness in processing time, reducing costs, and improving accuracy [30]. Calculation of the volume of fluid resuscitation is important especially for those at risk for complications such as heart failure. AI tools have been developed to accurately predict fluid requirements in the emergency department and intensive care unit. AI offers unique applications for detecting adverse events in the clinical trial phases of drug and medical device development, potentially reducing significant costs and preventing patient harm.

Diagnostics

Accurate and timely diagnosis is a multifactorial process that typically involves an interplay between system-related and cognitive factors of knowledge, experience, and effective patient-provider communication as well as available clinical and technical resources, rendering it high risk for errors. Diagnostic errors often occur over time from initial assessment to follow-up and arise when a diagnosis is incorrect, significantly delayed, or missed. Most people are likely to experience some diagnostic error throughout their life span. It is estimated that 5% of adults in high-income countries suffer from diagnostic errors each year in outpatient settings and the

magnitude is likely to be higher in low- and middle-income countries due to a paucity of healthcare professionals, limited testing resources, and record-keeping systems [31].

The widespread adoption of EHR presents an opportunity to automate tracking of diagnostic errors, understand cognitive biases, and assist healthcare systems with diagnoses. Deep learning algorithms, for example, showed a better diagnostic performance than a panel of 11 pathologists for the detection of lymph node metastases in women with breast cancer [32]. Similarly, EHR-based detection of familial hypercholesterolemia (FH) through random forest classifiers was found to identify thousands of undiagnosed patients with FH leading to better clinical management and screening of family members [33].

However, in contrast to structured datasets, a substantial amount of clinical narratives is semi- or unstructured which requires NLP and automated information extraction tools to extract information through mapping unstructured text into a semi-structured (or structured) form. Therefore, harnessing embedded clinical information through data mining techniques and concept recognition provides a far richer source of safety-related events and cognitive errors as well as reducing expensive labor-intensive chart reviews [34]. Only a handful studies have used unstructured textual data to identify diagnostic errors. Murff et al. (2011) showed that NLP-based approach had higher sensitivity and lower specificity in detecting postoperative complications compared to patient safety indicators of administrative discharge coding [35]. This is in line with previous published studies on the superior role of NLP in identifying genomic mutation-associated cancer treatment change and detecting safety-related events [36].

Medical imaging is another promising area for the application of AI and has the potential not only to improve image quality and diagnostic efficiency but also to optimize staffing and decrease radiation dosing through reducing scanning time. Once an image is captured, it can be incorporated alongside EHR and clinical narratives to aid diagnosis. A recent study illustrated the potential role for an AI system for the rapid identification of COVID-19 patients, combining CT imaging with

clinical information, showing a similar accuracy to a senior chest radiologist [37].

Despite the advances in big data analytics and their potential to improve diagnosis, the challenge remains in measuring the incidence of diagnostic errors. The lack of a standard metric to delineate the scope of the problem makes it difficult to gauge the effectiveness of AI techniques. All these safety issues and technological and political transformations working progressively in tandem will prosper in the creation of a data-driven system that reduces the global burden of diagnostic errors.

Medication Errors and Polypharmacy

Medication errors are a leading cause of injury and preventable harm in healthcare systems, and in the United States, it is estimated there is 1 medication error per hospitalized patient per day and is responsible for about 7000 deaths per year in hospitals [38]. Globally, the annual cost associated with medication errors has been estimated at US\$ 42 billion [1].

In developing and transitional countries, a report from the World Health Organization reviewing 679 studies from 97 countries found that over 50% of all medicines are prescribed, dispensed, or sold inappropriately and half of all patients failed to take medicines correctly and less than 30–40% of patients are treated according to some existing clinical guidelines [39]. With regard to quality of care in low-income countries, one of the very first papers analyzing 15,000 patient records hypothesized that poor quality of care is due to lack of medications, specialists, laboratory services, and other technology [40]. However, their findings showed that the two key factors causing adverse events were treatment errors and diagnostic errors, highlighting the central role of data in optimizing decision-making and therefore patient care [40].

Drug safety has been thought as the most significant contributor to overall medical errors and a target area for AI experts in this field. The Joint Commission on Accreditation of Healthcare Organizations highlighted that improper administration of medications or dosages are two key

sources of medication errors that result in severe patient harm and death [41]. Choudhury and Asan [42] reviewed studies exploring the influence of AI on safety outcomes and found that the majority of AI studies improved patient safety by preventing incorrect administration of drugs and overdoses and identifying adverse drug reactions.

Polypharmacy is a major global health issue posing a significant impact on safety outcomes and increases the probability of adverse medication events. Polypharmacy is often defined as the concurrent use of five or more medications including over-the-counter, prescribed, traditional, and complementary medicines [43]. Inappropriate polypharmacy, including nonadherence and socioeconomic factors, is estimated to contribute to 4% of total preventable costs worldwide, totalling USD 18 billion [44]. Gillespie et al. [45] performed a randomized-controlled trial that showed medication review by clinical pharmacists resulted in significant reduction in hospital visits, morbidity, and financial cost, a strategy that may enhance hospital-level safety measures.

There are a few studies to date which have investigated the role of AI techniques for predicting polypharmacy adverse events including DeepMind's acute kidney injury prediction, [46] a condition that may result from drug-drug interaction. Kocbek et al. [47] built balanced ML models of good performance and interpretability to forecast polypharmacy complications for patients suffering from either cardiovascular disease or type 2 diabetes, using drug prescription and hospital discharge datasets. A handful of studies implemented NLP on a small scale to predict polypharmacy harms or errors, highlighting the need to integrate ML and NLP models to identify high-risk patients and gain insights into the key variables for risk prevention [48]. The development of AI-driven decision support systems has the promise to identify appropriateness of polypharmacy and improve prescribing of medications.

Treatment Outcomes and Quality

Prevalent diseases and conditions in older persons in developed countries include brain diseases, osteoarthritis, hypertension, vision and hearing

loss, and diabetes. However, multimorbidity is common and increases the risk of complications due to treatment. Much of the current research using AI in the diagnosis and treatment of diseases are focused on single diseases or conditions, although risk stratification of patients with multimorbidity is being rapidly developed [49].

AI research in brain diseases is perhaps more advanced than other prevalent diseases. A systematic review [50] on the use of AI in brain diseases noted that the large amounts of data captured as medical imaging, free text, and data from monitoring devices have radically changed the identification and treatment of brain pathologies and tracking of the response to treatment [51]. Machine learning has particularly become widely used for conditions such as Alzheimer's disease, dementia, multiple sclerosis, and degenerative diseases. The use of AI in brain diseases is being applied to predict mortality, surgery response and postoperative hospitalization, and disease recurrence in a range of brain diseases and conditions [51].

Osteoarthritis, another common disease that is heavily reliant on imaging for diagnosis and response to treatment, is another area using AI. Deep learning has been useful in the diagnosis and assessment of knee and hip osteoarthritis and to assess knee pain [52]. However, although AI and deep learning are successfully being applied for diagnostic purposes, the use of AI in clinical trials remains nascent.

Hypertension is another prevalent disease and is the leading risk factor for cardiovascular mortality and morbidity. Despite being widespread and prevalent, treatment of hypertension has not changed significantly in the past two decades [53]. The two broad approaches are lifestyle modifications and pharmacotherapy. The role of AI in the treatment of hypertension has been proposed with the use of cuffless monitors and standardized electronic counselling through digital applications and pharmacogenetics-based algorithms [53].

Related to hypertension management is the management of age-related macular degeneration (AMD) that is the leading cause of blindness in older adults in developed countries. The incidence of AMD is expected to increase 1.5 times over the next 10 years due to an aging population and

increasing rates of hypertension [54]. Using machine learning with AMD-specific image parameter algorithms has shown promising results to predict progression of late AMD within 1–2 years.

Orthopedic surgery is heading in the direction of safety with the implementation of robotics to deliver joint replacement. Robots were first introduced 30 years ago, to abolish surgeon error, but only recently has their assistance been shown to be beneficial. The lack of objective metrics of joint health and disease and of prospective outcome data for many surgical treatments allows surgeons' selection and treatment biases to persist. This coupled with the notoriously poor accuracy of surgeons in predicting patient outcomes [55] and the lack of consensus on optimal management strategies, especially for common conditions, prompts the need for large unbiased data and the use of targeted AI systems. ML-derived prediction techniques are capable of aiding orthopods shift away from solely relying on clinical risk stratification models in making shared informed decisions with their patients [56]. For example, forecasting the probability of postoperative delirium and discharge destination following hip fractures has enhanced efficiency and safety and quality measures while providing opportunities for preventive targeted interventions [57, 58].

Clearly the use of AI in the treatment and outcomes of prevalent medical and surgical diseases is advancing quickly. There is enormous potential for the use of AI in the diagnosis, treatment, and management of multimorbidity, particularly in polypharmacy, which puts patients at greater risk of complications [59].

Patient Safety Databases

By now, we hope it is clear to the readers that the biggest challenge to advancing AI in the area of patient safety is the lack of large, comprehensive, and high-resolution clinical data.

The Centers for Disease Control and Prevention (CDC) and Centers for Medicare and Medicaid Services (CMS) compile reports of hospital-acquired infections, and the FDA collects data on adverse drug events, but none of these

datasets are comprehensive and high-resolution in order to build prediction, classification, or optimization models. There are a number of "opt-in" databases for medical errors, such as the Data Sharing Project (DSP) which was launched in 1985 as an independent collaborative effort to have a solid infrastructure of medical professional liability (MPL) claims and lawsuits. Given that these are mainly a collection of anonymous reports, they cannot be linked with other databases, and are therefore also not particularly valuable for machine learning research.

In New Zealand, the Accident Compensation Corporation (ACC) was established in 1974 and has provided compulsory insurance cover for treatment injuries for everyone in New Zealand since 2005 [60]. The dataset is a comprehensive nationwide reporting of medical errors in a no-fault compensation healthcare system. The dataset is the most comprehensive reliable and unbiased dataset for error and injury claims worldwide.

Future of AI in Safety and Quality

Although the advent of AI offers paths toward transforming the way safety and quality are delivered, an emphasis on their generalizability and interpretability is paramount. To create AI systems that are clinically useful, healthcare professionals should develop an understanding of how, when, and why these systems work, shifting away from the traditional geographical generalizability of clinical research [61]. AI systems are a set of algorithms that were trained to learn and operate under specific assumptions and clinical contexts, lacking the adaptability of expert clinicians. This prompts the need to replace our narrow focus on generalizability with a broader construct of the ultimate goal of AI at the bedside.

At present, AI techniques are usually evaluated by their precision and accuracy, but just because a system is deemed accurate, it does not imply unbiased or fair patient outcomes. Several key ethical considerations have been identified regarding the use of AI in safety and quality research, illuminating racial disparities. The AI community must engage in a holistic perspective in

addressing any disparities and to ensure that such systems will not perpetuate or magnify any marginalized communities' biases. A model pipeline framework is considered as a stepping-stone for AI healthcare ethics in addressing structural inequities, but the responsibility always lies on us as a society to ensure fairness and transparency.

Above all, the challenge in utilizing AI to optimize patient safety and quality remains in accessing a reliable and unbiased database. The ACC dataset provides a solid platform to address many of the global safety challenges considering New Zealand's unique no-fault compensation healthcare system. With the scale of patient impacts at stake and through working with a diverse team of physicians, developers, and engineers while closely understanding and inspecting potential biases and subpopulation disparities, we envision ACC to become a center of excellence in safety and quality science.

Using AI technologies may help create a transparent predictive platform that improves the shared decision-making process among patients, doctors, and policy makers. For example, an AI system would be a valuable tool to detect and track treatment injuries (TI), set benchmarks of medical errors, and may assist policy makers on establishing TI-related guidelines as well as identify TI-related factors. Understanding these predictors is arguably believed to improve the performance of healthcare systems worldwide. Determining which patients will benefit most from which form of medical intervention while understanding associated risks may also help inform policy, minimizing risk and maximizing health gain.

References

- World Health Organization. Regional strategy for patient safety in the WHO South-East Asia Region (2016–2025). WHO Regional Office for South-East Asia; 2015.
- Grober ED, Bohnen JMA. Defining medical error. *Can J Surg*. 2005;48(1):39–44.
- Wallis KA. Learning from no-fault treatment injury claims to improve the safety of older patients. *Ann Fam Med*. 2015;13(5):472–4.
- Shojania KG, Duncan BW, McDonald KM, Wachter RM, Markowitz AJ. Making Healthcare Safer: a critical analysis of patient safety practices. Rockville: Agency for Healthcare Research and Quality; 2001. Evidence Report/Technology Assessment No. 43; AHRQ publication 01-E058.
- Newman-Toker DE, Pronovost PJ. Diagnostic errors—the next frontier for patient safety. *JAMA*. 2009;301(10):1060–2.
- Macrae C. Governing the safety of artificial intelligence in healthcare. *BMJ Qual Saf*. 2019;28(6):495–8.
- Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med*. 2018;24:1716–20.
- McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577:89–94.
- Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25:44–56.
- Tanner C, Gans D, White J, Nath R, Pohl J. Electronic health records and patient safety: co-occurrence of early EHR implementation with patient safety practices in primary care settings. *Appl Clin Inform*. 2015;6(1):136–47.
- ACS NSQIP [Internet]. American College of Surgeons. 2020 [cited 16 November 2020]. <https://www.facs.org/quality-programs/acs-nsqip/about>
- Neves A, Freise L, Laranjo L, Carter A, Darzi A, Mayer E. Impact of providing patients access to electronic health records on quality and safety of care: a systematic review and meta-analysis. *BMJ Qual Saf*. 2020;29:1019–32.
- Bell SK, Delbanco T, Elmore J, Fitzgerald P, Fossa A, Harcourt K, et al. Frequency and types of patient-reported errors in electronic health record ambulatory care notes. *JAMA Netw Open*; 2020; 3(6).
- Parkinson B, Meacock R, Checkland K, Sutton M. How sensitive are avoidable emergency department attendances to primary care quality? Retrospective observational study. *BMJ Qual Saf*. 2020;0:1–9.
- Leape LL. Reporting of adverse events. *N Engl J Med*. 2002;347(20):1633–8.
- ANZACS-QI (Registry) | Enigma Solutions Ltd [Internet]. Enigma.co.nz. 2020 [cited 16 November 2020]. <https://www.enigma.co.nz/predict-medical/anzacs-q/>
- Kerr AJ, Lee M, Jiang Y, Grey C, Wells S, Williams M, Jackson R, Poppe K. High level of capture of coronary intervention and associated acute coronary syndromes in the all New Zealand acute coronary syndrome quality improvement cardiac registry and excellent agreement with national administrative datasets (ANZACS-QI 25). *N Z Med J*. 2019;132(1492):19–29.
- Stewart R. Comparison of high and low oxygen protocol in patients with suspected ACS. Presentation presented at; 2019; ESC Congress.
- My audits [Internet]. 2020 [cited 16 November 2020]. <https://www.surgeons.org/en/research-audit/my-audits>
- Finlayson SR. Assessing and improving the quality of surgical care in rural America. *Surg Clin North Am*. 2009;89(6):1373–81.

21. Anderson KT, Appelbaum R, Bartz-Kurycki MA, Tsao K, Browne M. Advances in perioperative quality and safety. *Semin Pediatr Surg.* 2018;27(2):92–101.
22. Reason J. Human error: models and management. *BMJ.* 2000;320(7237):768–70.
23. Braithwaite J, Donaldson L. Patient safety and quality. In: Ferlie E, Montgomery K, Pederson AR, editors. *The Oxford handbook of health care management.* Oxford, UK: Oxford University Press; 2016. p. 325–51.
24. Braithwaite J, Wears RL, Hollnagel E. Resilient health care: turning patient safety on its head. *Int J Qual Health Care.* 2015;27(5):418–20.
25. McCarthy J, Hayes P. Some philosophical problems from the standpoint of artificial intelligence. In: Meltzer B, Michie D, editors. *Machine Intelligence 4.* Edinburgh: Edinburgh University Press; 1969. p. 463–502.
26. World Health Organization. World alliance for patient safety: WHO draft guidelines for adverse event reporting and learning systems: from information to action. World Health Organization; 2005. <https://apps.who.int/iris/handle/10665/69797>
27. Report by the Director General – A72/26. Patient safety – global action on patient safety. Geneva: World Health Organization; 2019. https://apps.who.int/gb/ebwha/pdf_files/WHA72/A72_26-en.pdf. Accessed 20 Nov 2020.
28. Schwendimann R, Blatter C, Dhaini S, Simon M, Ausserhofer D. The occurrence, types, consequences and preventability of in-hospital adverse events – a scoping review. *BMC Health Serv Res.* 2018;18:521.
29. Unbeck M, Schildmeijer K, Henriksson P, Jürgensen U, Muren O, Nilsson L, Pukk-Härenstam K. Is detection of adverse events affected by record review methodology? An evaluation of the “Harvard Medical Practice Study” method and the “Global Trigger Tool”. *Patient Saf Surg.* 2013;7(1):10.
30. Schmider J, Kumar K, LaForest C, Swankoski B, Naim K, Caubel PM. Innovation in pharmacovigilance: use of artificial intelligence in adverse event case processing. *Clin Pharmacol Ther.* 2019;105(4):954–61.
31. Singh H, Schiff GD, Gruber ML, Onakpoya I, Thompson MJ. The global burden of diagnostic errors in primary care. *BMJ Qual Saf.* 2016;0:1–11.
32. Bejnordi BE, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA.* 2017;318(22):2199–210.
33. Banda JM, Sarraju A, Abbasi F, et al. Finding missed cases of familial hypercholesterolemia in health systems using machine learning. *NPJ Digit Med.* 2019;2(23):1–8.
34. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform.* 2008;17(1):128–44.
35. Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, Dittus RS, Rosen AK, Elkin PL, Brown SH, Speroff T. Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA.* 2011;306(8):848–55.
36. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. *J Am Med Inform Assoc.* 2005;12(4):448–57.
37. Mei X, Lee HC, Diao KY, et al. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. *Nat Med.* 2020;26:1224–8.
38. Institute of Medicine (US) Committee on Quality of Health Care in America; Kohn LT, Corrigan JM, Donaldson MS, editors. *To Err is Human: Building a Safer Health System.* Washington, DC: National Academies Press (US); 2000; 2, Errors in health care: a leading cause of death and injury. <https://www.ncbi.nlm.nih.gov/books/NBK225187/>
39. Shankar PR. Medicines use in primary care in developing and transitional countries: fact book summarizing results from studies reported between 1990 and 2006. *Bull World Health Organ.* 2009;87(10):804.
40. Wilson RM, Michel P, Olsen S, et al. Patient safety in developing countries: retrospective estimation of scale and nature of harm to patients in hospital. *BMJ.* 2012;344:1–14.
41. Barnsteiner JH. Medication reconciliation. In: Hughes RG, editor. *Patient safety and quality: an evidence-based handbook for nurses.* Rockville: Agency for Healthcare Research and Quality (US); 2008. Chapter 38. <https://www.ncbi.nlm.nih.gov/books/NBK2648/>.
42. Choudhury A, Asan O. Human factors: bridging artificial intelligence and patient safety. *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care.* 2020;9(1):211–15.
43. Thomson W, Farrell B. Deprescribing: what is it and what does the evidence tell us? *Can J Hosp Pharm.* 2013;66(3):201–2.
44. Aitken M, Gorokhovich L. Advancing the responsible use of medicines: applying levers for change. IMS Institute for Healthcare Informatics: Parsippany; 2012.
45. Gillespie U, Alasaad A, Henrohn D, Garmo H, Hammarlund-Udenaes M, Toss H, et al. A comprehensive pharmacist intervention to reduce morbidity in patients 80 years or older: a randomized controlled trial. *Arch Intern Med.* 2009;169(9):894–900.
46. Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature.* 2019;572:116–9.
47. Kocbek S, Kocbek P, Stozer A, Zupanic T, Groza T, Stiglic G. Building interpretable models for polypharmacy prediction in older chronic patients based on drug prescription records. *PeerJ.* 2018;6:e5765.
48. Wilfling D, Hinz A, Steinhäuser J. Big data analysis techniques to address polypharmacy in patients – a scoping review. *BMC Fam Pract.* 2020;21:180.

49. Obermeyer Z, Emanuel EJ. Predicting the future – big data, machine learning, and clinical medicine. *New Engl J Med.* 2016;375(13):1216–9.
50. Segato A, Marzullo A, Calimeri F, Elena De Momi E. Artificial intelligence for brain diseases: a systematic review. *APL Bioeng.* 2020;4(4):041503.
51. Senders JT, Arnaout O, Karhade AV, Dasenbrook HH, Gormley WB, Broekman ML, Smith TR. Natural and artificial intelligence in neurosurgery: a systematic review. *Neurosurgery.* 2018;83:181–92.
52. Hayashi D, Roemer FW, Eckstein F, Samuels J, Guermazi A. Imaging of OA – from disease modification to clinical utility. *Best Pract Res Clin Rheumatol.* 2020;101588.
53. Kulkarni S. Hypertension management in 2030: a kaleidoscopic view. *J Hum Hypertens.* 2020;1–6.
54. Wong TY, Liew G, Mitchell P. Clinical update: new treatments for age-related macular degeneration. *Lancet.* 2007;370:194–206.
55. Bloembergen CH, van de Graaf VA, Virgile A, et al. Infographic: can even experienced orthopaedic surgeons predict who will benefit from surgery when patients present with degenerative meniscal tears? A survey of 194 orthopaedic surgeons who made 3880 predictions. *Br J Sports Med.* 2020;54(9):556–7.
56. Oosterhoff JHF, Doornberg JN. Artificial intelligence in orthopaedics: false hope or not? A narrative review along the line of Gartner's hype cycle. *EFFORT Open Rev.* 2020;5:593–603.
57. Oosterhoff JHF, Karhade AV, Oberai T, Doornberg JN, Schwab JH. Development of machine learning algorithms for prediction of postoperative delirium in elderly hip fracture patients. Manuscript submitted for publication to *Clin Orthop Relat Res*; 2019.
58. Pitzul KB, Wodchis WP, Kreder HJ, Carter MW, Jaglal SB. Discharge destination following hip fracture: comparative effectiveness and cost analyses. *Arch Osteoporos.* 2017;12:87.
59. Molokhia M, Majeed A. Current and future perspectives on the management of polypharmacy. *BMC Fam Pract.* 2017;18(1):70.
60. Accident Compensation Act, Stat. 49, 2001 (NZ). http://www.legislation.govt.nz/act/public/2001/0049/latest/DLM99494.html?search=ts_act_accident_resel&p=1&sr=1
61. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Heal.* 2020;2:489–92.



Right to Contest AI Diagnostics

15

Defining Transparency and Explainability Requirements from a Patient's Perspective

Thomas Ploug and Søren Holm

Contents

Introduction	228
The Right to Effective Contestability and Transparency	229
The Four Dimensions of Contestability	231
Dimension 1: Personal Health Data	232
Dimension 2: Bias	232
Dimension 3: Performance	234
Dimension 4: Decisional Role	235
Further Issues and Aspects of the Right to Contest	236
References	237

Abstract

The problem of the transparency and explainability of AI decision-making has attracted considerable attention in recent years. In this chapter, we argue that patients have a right to contest AI medical decisions and that the transparency requirements of AI

decision-making in health care should be guided by this right. We define the right to contest AI medical decisions both formally and substantially. Formally, the right to contest AI medical decisions must be a right (i) that is grounded in moral values, (ii) that is effective in protecting patients' rights, (iii) that is proportional to the potential costs to others, and (iv) that is an application of a more general right to contest medical decisions. Substantially, the right to contest AI medical decisions should enable patients to contest (I) the AI system's use of personal and sensitive data, (II) the system's potential biases, (III) the system performance, and (IV) the division of labor between the system and healthcare professionals. We justify and define 14 specific informational requirements – i.e., transparency requirements – that follow from the substantial notion of the right to contest AI medical

T. Ploug (✉)
Centre for Applied Ethics and Philosophy of Science,
Department of Communication, Aalborg University
Copenhagen, Copenhagen, Denmark
e-mail: ploug@hum.aau.dk

S. Holm
Centre for Social Ethics and Policy, School of Law,
University of Manchester, Manchester, UK

Center for Medical Ethics, Faculty of Medicine, University
of Oslo, Oslo, Norway
e-mail: soren.holm@manchester.ac.uk

decisions. Finally, we briefly discuss the patient-centered approach taken in this chapter against alternative approaches grounding transparency requirements in considerations of democracy and the interests of science.

Keywords

Contestability · Transparency · Explainability · Privacy · Bias · Performance · GDPR · Artificial intelligence · Medical decision-making

Introduction

Imagine two closely knit families having spent their holiday together. Celebrating their friendship and the memorable holiday, they go for a large dinner of authentic local foods on the last night of their holiday. Upon their return an adult member of each of the families, say Sarah and Bob, develops very similar, quite unpleasant symptoms of gastroenteritis. Sarah seeks healthcare assistance in the local clinic. The clinic has recently introduced the latest machine learning-based diagnostic and treatment planning support system – “Dr A.I. Grey 5.0.” Based on system recommendations, Sarah is further examined, diagnosed with amoebic dysentery, and prescribed metronidazole. Learning about Sarah’s diagnosis, Bob also turns to the local clinic for diagnosis and treatment. The system concludes that the gastrointestinal symptoms are indicative of a mild case of irritable bowel syndrome and will pass. Bob is recommended rest and fiber supplements to avoid constipation, although unsettled and inquisitive Bob is offered no further explanation or consultation.

The use of artificial intelligence decision-support systems in health care holds great potential for improved and more cost-efficient diagnostics and treatment planning. A recent review found that in terms of diagnostic performance deep learning models may outperform healthcare professionals in a number of cases [1]. The review also showed that in general the diagnostic performance of deep learning

models is equivalent to that of healthcare professionals, although the research providing this evidence is often of low quality. However, the potential gains in improved and more cost-efficient diagnostics come at a cost. Deep learning models are opaque. The reasoning procedure of complex machine learning systems – and deep learning systems in particular – cannot be fully explicated and replicated by a human being [2–7]. There is widespread consensus that the transparency of machine learning systems is desirable. A recent systematic review identified 84 guidelines on responsible use of AI issued by various different stakeholders [8]. The guidelines all emphasize the importance of transparency but differ in their definition of transparency and its requirements. The focus on transparency is also driven by the legal requirements in the EU’s General Data Protection Regulation which states in Article 13 that the data controller must “...provide the data subject with the following further information necessary to ensure fair and transparent processing [...] the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, **meaningful information about the logic involved**, as well as the significance and the envisaged consequences of such processing for the data subject” (our emphasis) [9]. There is an identical clause in Article 14, and the requirement for “meaningful information about the logic involved” has in the literature been interpreted as a requirement for transparency, although the exact scope of the requirement is contested [10–12].

There is need of identifying a single guiding principle or reason behind a notion of transparency and in particular behind a notion of transparency of machine learning-based decisions in the healthcare context. There are several candidates. Thus, the transparency of machine learning-based diagnostics and treatment planning is important for several reasons. Transparency is important for scientific reasons. Machine learning models establish statistical correlations that may be suggestive or perhaps even constitutive of causal mechanisms [6, 13]. Identifying causal mechanism is key to understanding diseases and

initiating preventive or therapeutic interventions. Transparency is also important for democratic reasons. In a public healthcare setting, the use of machine learning models for diagnostic and treatment planning purposes is a way in which the State exercises power over its citizens, and it is a way of distributing healthcare goods. As anywhere else where the State exercises power over its citizens and makes distributive decisions, transparency should be sought in order for the citizens to be able to hold the State and the decision-makers accountable through the institutions of democracy. In the healthcare system, this is the basis for one of the most widely used underpinning frameworks for priority setting, Accountability for Reasonableness [14].

The simple case of Sarah and Bob illustrates what we believe to be the most fundamental ethical reason for insisting on the transparency of AI decision-making. The AI-generated diagnosis of Bob may or may not be wrong, but what is certainly wrong here is that Bob is in effect robbed of the opportunity to protect himself by contesting the AI diagnosis and treatment suggestion. Essentially, this chapter argues that patients have a right to contest the AI medical decisions and that the transparency of AI decision-making in health care should be guided by this right. We furthermore argue that this right has four dimensions together entailing that 14 types of information about an AI system must be accessible by a patient. Finally, we consider how contestability is linked to some of the other reasons for insisting on the transparency of AI decision-making. All throughout, the term “AI” is taken to denote complex machine learning systems.

The Right to Effective Contestability and Transparency

The right to contest AI decisions advocated in this chapter is a fundamental right to protect oneself by objecting to AI-generated decisions that one believes may have adverse effects on one’s life. It is not only a right to object to these effects but also to have these objections evaluated and potentially rectified.

The right to contest AI decisions presupposes and entails transparency but also at the same time provides an explication of and justification for what kind of transparency we should aim at. If Bob has a right to contest the AI-generated diagnosis and treatment plan, then he must have a right to be provided with some information about the involvement of AI in the diagnostic process. Without access to information, Bob cannot contest the AI decisions in a meaningful way. The real challenge here is therefore to determine the scope and limits of this right to contest and the entailed right to information. We believe that an adequate notion of the right to contest AI decision in health care should satisfy the following four conditions; it must be as follows:

- Value-based
- Effective
- Proportional
- General

That the right to contest AI decisions requires further delineation is readily seen. Bob may have interests giving him reason to contest all sorts of things about AI use in health care. He may want to contest the use of AI systems produced in a particular country or by people of a certain ethnicity or with a certain hair color, but this cannot reasonably be thought to provide him with a right to be listened to nor to information about any of these things. Reaching this conclusion requires, however, a more comprehensive framework for defining positively the content and scope of the right to contest.

We believe that *moral values* should determine the limits of the right to contest and the entailed right to information. That is, the right to contest should be defined and limited by the set of moral values and implied moral rights that it is intended to protect. To illustrate, if Bob’s right to contest is partly grounded in the value of protecting Bob against any physical harm suffered as a consequence of, e.g., inaccurate AI diagnostics and treatment planning – and Bob’s right to protect himself against such harm – then this explains both why Bob has a right to contest the AI-generated diagnosis and treatment plan and

why he does not have the right to contest the use of AI because of the hair color of its developers. Accordingly, Bob should be offered information of relevance for contesting wrongful AI diagnosis and treatment, and not information about the hair color of the developers. Taking this value-based approach implies that any claim to more information must start by showing why this information is relevant for the protection of individuals and their rights. It must make reference to some values and rights and explain why access to a certain type of information is required in order to determine if these values and rights have been or may be violated. Note that, while moral values may determine what information is *relevant* for a right to information, it does not determine the specific character of this information in every detail. Information about the accuracy of AI diagnostics may, for instance, be provided in many different formats.

Secondly, we believe that the right to contest AI decisions must be a right to *effective* contestation. That is, it must be a right to make objections that are as effective as possible in protecting those values underlying the very right to object. If Bob's right to contest is, as suggested above, partly rooted in his right to protect himself against the harms of inaccurate AI diagnostics and treatment planning, then it must be a right to be able to do so as effectively as possible. It must be a right that enables him to contest that a specific AI system with certain algorithms and trained on a particular dataset is being used in a particular organizational setup that seems to have generated an inaccurate diagnosis and treatment. The use of medical technology embedded with machine learning, e.g., for recording and interpreting ECGs, does not in and of itself cause harm to Bob – it is likely to his benefit. However, if Bob is offered no other option, he may choose to contest the very use of such medical technology as part of the diagnostic process. Contesting all forms of AI use in clinical context would be a very counterproductive way of protecting himself against harm since it may end up having harmful consequences for him. Contesting specific aspects of the AI system used for diagnostics and treatment planning in the specific organizational setup would be more

efficient. This delineation of the right to effective contestation has implications for the information Bob should have access to. He cannot most efficiently protect himself on the basis of general and abstract information about the AI use in health care, e.g., "AI technology is being used in this clinic." In order to effectively protect himself, he must have access to information about specific features of the AI system used and the organizational setup that applies to his particular case.

Thirdly, the right to contest AI decision-making in health care must be proportional. That is, it must balance proportionally individuals' right to protect themselves against violations of their rights due to AI system involvement in the diagnostic process against the potential benefits for others and society of such use. The right to contest AI decision-making in health care cannot in and of itself justifiably entail the impossibility of developing and implementing AI in health care that may be beneficial for others and society. The right to contest is a right to self-protection. It is not a right that in and of itself justifies limiting disproportionately the access of others to significant benefits. There may well be other moral reasons for limiting the use of AI in health care. The point here is that a right to contest AI use cannot justifiably constitute such a reason. It follows that the right to contest cannot justifiably entail transparency requirements – i.e., a right to information – that effectively makes impossible the very development and use of AI in health care. However, giving patients access to certain types of information that enables them to contest effectively the involvement of AI in the diagnostic process will inevitably have costs for both for system developers and clinicians that will have to retrieve and make this information accessible. As we aim to show in the following section, this information does, however, not have to be of a character that imposes disproportionate costs on the developers and healthcare professionals.

Fourth and finally, the right to contest AI decision-making in health care should be an application of a *general right* to contest medical decisions and correspondingly a broader right to gain access to information about the medical decisions. It cannot be a right introduced exclusively in

relation to AI for the simple reason that there is no ethically relevant difference that would justify such narrowing of the scope of the right to contest. If Bob had received his diagnosis and treatment plan by a doctor, there would be no less reason to insist that he should have the right to contest the decision and that this right entails that he should have access to essential information about how and why the doctor arrived at the relevant diagnosis and treatment plan. Doctors and AI systems are alike in the sense that they may fail to provide adequate protection of patients and their rights, and they may fail to do so for reasons we may never – for both practical and principled reasons – be able to fully uncover and understand. Research into biases in human decision-making reveals a striking amount of such biases [15]. Therefore there cannot be double standards in relation to the right to information, i.e., transparency. There has to be only one answer to the question of what information patients should have access to: the information required in order to effectively contest medical decisions. However, what information that is certainly depends on who is making the medical decision. Not all questions can meaningfully be posed to both AI systems and doctors. In order to contest the accuracy of an AI diagnosis and treatment plan, it may be considered relevant to know something about the quality of the dataset on which the algorithm was initially trained. It does not make sense to make the same inquiry about doctors whom it would be more obvious to question about their special field of training, competence, and interest and their level of experience. On the other hand, very similar questions about conflict of interest arise and must be answered in relation to the choice of treatment and the commercial influences on that choice. Ultimately, the general right to information may thus in the AI context come to be a right to types of information defined by special features of AI systems.

Satisfying the four conditions laid out above does not by definition entail that all aspects of AI decision-making must be explicable and/or replicable by a human being. They do not by definition require that we have to be able to fully open the “black box.” On the contrary, they only require that patients should have access to *sufficient*,

relevant, and specific information that enables them to *effectively* and *proportionally* contest medical decisions that may harm them or in some other way violate their rights. And that is wholly consistent with current practices in health care which we believe to be morally justified. Consider, for example, the use of medication. Drugs are often also “black boxes” in that the exact mechanism of action – the exact causal mechanism – is partly or wholly unknown. We are still learning new things about exactly how aspirin, on the market since 1888, works. Treating patients with a particular drug is, however, contestable in the sense that the patients can be provided with sufficient, relevant, and specific information about the basis for this decision, including information about the likelihood of effects and side effects, to be able to protect themselves against harm and other rights violations by exercising their right to object. Knowledge about the mechanism of action of a given drug may facilitate a better prediction of the effects and side effects, but information about the effects and side effects could also gradually become available through standard methods of testing. When a drug has been made available, the information required for effective contestation will also be available.

The Four Dimensions of Contestability

Contesting AI involvement in diagnostics and treatment planning should encompass four different aspects of this involvement and the AI system:

1. The use of personal health data by AI
2. The potential bias of AI
3. The performance of AI
4. The decisional role of AI

In what follows we shall establish more firmly the relevance of these dimensions, show how they are grounded in basic moral values and individual moral rights, and finally provide a detailed set of information requirements implied by the patients’ right to contest the AI involvement. In so doing, we will have provided a substantial notion not only of

the right to contest AI involvement in medical decisions but also of the corresponding right to the transparency of such AI involvement. For each of the dimension, we will develop specific contestability variables, leading to contestability questions and required contestability information. These are summarized in Table 1. The content of this section builds on our previous work on this topic [16].

Dimension 1: Personal Health Data

The use of AI models and systems for diagnostic and treatment planning purposes presupposes access to personal health data. Bob's diagnosis and treatment plan is the outcome of algorithms being applied to data of relevance for his present condition, e.g., clinical tests, personal health record data, scans, tissue, data from wearables, and non-health data collected digitally by various actors. Personal health data is sensitive for a number of reasons [17], and studies indicate that a majority of people are concerned about the privacy of their health record or the exchange of health data in the healthcare system [18–20]. The AI system use of personal health data may thus be considered invasions of privacy in the sense of involving the exchange and use of personal and sensitive data. However, personal health data may also be misleading by virtue of being outdated, one-sided, erroneous, and incomplete [21]. The AI system use of personal health data may consequently lead to misdiagnosis and inadequate treatment and thus ultimately come to have harmful consequences.

The potential of privacy invasions and harm makes relevant the right to privacy and the right not to be harmed, and not least the corresponding rights to protect oneself against violations of these rights. We believe that individuals should have a right to protect themselves against violations of privacy and harm by contesting the use of sensitive health data and other personal data in AI diagnostics and treatment planning. In order to be able to exercise this right efficiently, they should be provided sufficient, relevant, and specific information about the use of their personal

and sensitive health data. What information is that? Minimally, patients should be able to contest the use of health and personal data on grounds of its sensitivity and quality. This may be taken to imply that patients minimally should have access to information about (1) the types of personal data used by the AI system, e.g., clinical test data, genetic data, life style data etc., and (2) the sources of this data, e.g., electronic patient record, wearables, etc. The combination of information about both types and sources of data provides individuals with a fair chance of assessing the sensitivity and quality of the data used and to contest accordingly.

Dimension 2: Bias

AI decisions in health care may be biased [22–29]. Biased decisions may tentatively be defined as decisions that systematically differentiates between groups of people beyond what is warranted. What is warranted may refer to at least two different types of baselines. It may refer to what is warranted by the evidence in which case a biased decision is a decision that differentiates between groups people in ways that do not reflect the evidence. An AI system recommending painkillers on the basis of previous prescriptions by doctors could systematically recommend different painkillers for young and old or black and white. In so far as this pattern in the prescriptions would not be found in a representative set of prescriptions, it would not be a recommendation warranted by the evidence. The bias could have various different sources, e.g., the AI system may have been trained on a non-representative dataset, the dataset could contain systematic errors in the human tagging, or the algorithm could contain an explicit bias. It is well recognized that financial conflicts of interest can cause bias in medical research and medical decision-making [30, 31]. There is no reason to believe that the same could not be the case in relation to the development of AI systems. If the development of an AI system in a particular therapeutic area was funded by a pharmaceutical firm with a large commercial interest in the area, there

Table 1 The Four Dimensions of Contestable AI and Associated Informational Requirements

Contestability – object	Contestability – variables	Contestability – questions	Contestability – required explanation
Data	Types of data	What types of personal data are used?	1) The decision D was made on the basis of data of type X, Y, Z about you
	Data sources	Where do your data come from?	2) The decision D was based on data from sources X, Y, Z
Bias	Training data	On which data was the AI trained?	3) The decision D was made by a system trained on existing data of type X, Y, Z
	“Tagging groups”	Who tagged training data?	4) The decision D was made by a system trained on data tagged by X, Y, Z
	Tested for bias	Was training data or AI system tested for bias?	5) The decision D was made by a system tested for bias of type X, Y, Z
	Conflict of interest	Is there a relevant financial conflict of interest?	6) There is no conflict of interest/there is a conflict of interest of type X, Y, Z
Performance	Accuracy	What is the accuracy of the AI system?	7) The decision D was made by a system with an accuracy of X, Y, Z
	Accuracy testing	How was the accuracy determined?	8) The decision D was made by a system with an accuracy determined by tests X, Y, Z
	Essential indicators	What are key variables of AI decision-making?	9) The key input data resulting in decision D was X, Y, Z
	Alternatives	Are alternatives considered?	10) The alternatives to decision D are X, Y, Z with a probability of x, y, z ($< d$)
	Longevity	When is decision	11) The decision D will be reconsidered if conditions X, Y, Z obtain
Decision	AI involvement	To what degree is AI making the decision?	12) The decision D involved an AI system with respect to X, Y, Z
	Human involvement	To what degree are humans making the decision?	13) The decision D was wholly/partly made by health professionals X, Y, Z
	Responsibility	Who is responsible for the decision?	14) The objective/legal responsibility for decision D is held by X, Y, Z

would be a conflict of interest and at the very least a potential for bias in the final system.

The baseline could also make reference to what is morally justified. In that case, a biased decision would be a decision that differentiates between groups of people in ways that constitute unfair or unjust differential treatment, i.e., differential treatment which violates fundamental moral values or rights. The AI system recommending different painkillers to young and old or black and white could be biased in this moral sense. If the painkillers are not relevantly similar – have the same effect, side effects, price, etc. – the difference in recommendation would arguably amount to de facto discrimination, i.e., it would violate basic principles of justice, fairness, and equality. Moral bias of AI decision-making may have all the same sources as statistical bias, but it may also result if all the data on doctors’ previous recommendations of

painkillers are reflective of a discriminatory practice. Statistical and moral bias are not thus co-extensional. If the AI system is trained on complete data reflecting a discriminatory practice in previous prescription of painkillers, then the difference in recommendations will be morally biased but statistically unbiased. Conversely, if there are no relevant differences in effects, side effects, price, etc. of the different painkillers, then a difference in recommendations – due to the AI system being trained on a non-representative dataset – will be statistically biased but morally unbiased.

Biased AI decision-making may not only violate principles of equality, fairness, and justice, it may also lead to misdiagnosis and inadequate treatment of the groups being discriminated against. Bob may not only be the victim of gender-based discrimination, but this discrimination is likely to result in harm following the wrong

diagnosis and the failure to receive adequate treatment as well as the harm resulting from the wrong treatment, i.e., by feeling compelled to eat excessive amounts of fibers. However, the potential bias of AI systems is not unavoidable. Prior to its implementation, an AI system may be tested for bias by applying the AI model to datasets differing on known “triggers” of discrimination, e.g., gender, age, ethnicity, and the like. Indications of bias may also come from testing an AI system against alternative systems or against doctors’ diagnostics and treatment planning.

The potential of AI bias resulting in discrimination and harm makes individuals’ right not to be discriminated against and not to be harmed relevant and the corresponding right to protect these rights. We believe that individuals should have a right to protect themselves against violations of privacy and harm by contesting the potential bias of AI systems. This requires sufficient, relevant, and specific information about the potential bias of the AI system. Minimally this should include information about (1) the character of the dataset on which the model is built, where by character is meant both the type and source of the training data. This is relevant for assessing the potential of bias inherent in data. For machine learning systems based on supervised learning, it also becomes relevant to know (2) who labeled the data and what instructions they were given concerning how to label. This is relevant for assessing the potential of bias occurring as a result of the labeling. For any type of machine learning system, it is important to know (3) the character and level of testing for bias that the AI system has undergone. Finally, it is important to know (4) whether there are any financial conflicts of interest in relation to the development of the AI system. Knowing that an AI system and the training data has been tested for bias and knowing about the quality of this testing is clearly important for assessing the potential of bias.

Dimension 3: Performance

As we have seen, AI diagnostic systems may issue a wrong diagnosis if applied to erroneous health

data. However, they may also – as in the case of biased systems – fail due to flaws in the underlying AI model. The previously referenced review concluded that the performance of deep learning models in health care is generally on a par with healthcare professionals [1]. The accuracy of the AI diagnostic model – including true and false positives and negatives – may be determined through tests in which the model is applied to sets of validated data of varying composition and size and for which the diagnosis is known. The accuracy may and may not change in the course of time. AI models that do not evolve over time, so-called “locked” AI models, have a set accuracy when applied to patient populations that are qualitatively identical to the population on which the model is trained [16]. While the accuracy as well as bias of evolving models may change, and they therefore seem unlikely to be introduced into routine healthcare practice, the accuracy could in principle be determined before the implementation and at set intervals thereafter.

The possibility of inaccurate diagnostics means that AI systems may cause harm by leading to over- or undertreatment. Bob may not only be harmed from inadequate treatment of what may be a case of amoebic dysentery, but he may also – albeit to minor degree – come to suffer from his eating of excessive amounts of fibers. The possibility of such harm invokes the patients’ right not to be harmed and their corresponding right to protect themselves. We believe that patients should be able to protect themselves against such harm by contesting the performance of AI diagnostic systems, and this entails that they should have access to information about (1) the accuracy of the AI diagnostic model and (2) the character and reliability of the tests conducted, including information about the time of the most recent tests in the case of evolving models. This information is, however, not sufficient, relevant, and specific enough to enable effective contestation of AI diagnostics. A doctor providing statistics on his previous diagnostic performance will not have provided information enabling a patient to effectively protect himself against the potential harm following the diagnosis the patient is faced with. Information about previous diagnostics does not

rule out the possibility that the diagnosis at hand is arrived at in some objectionable way likely to make it inaccurate, e.g., by tossing a coin. In order to effectively contest the diagnosis at hand – whether issued by an AI system or doctor – the patient must be provided with information about (3) the key indicators leading to that diagnosis, (4) alternatives to the suggested diagnosis, and (5) the changes in the patient’s condition that will lead to a reconsideration of the diagnosis. Information about key indicators enables the patient to form an opinion on the credibility of the suggested underlying pathology and the application of this pathology to the situation in question – and to contest accordingly. Information about alternative diagnoses and changes leading to a reconsideration of the diagnosis enables the patient to form an opinion on the quality of the diagnostic procedure – and to contest accordingly. Diagnoses are underdetermined by indicators in most cases, unless there is a definitive test or a pathognomonic sign or symptom, and the diagnostic procedure is best understood as a procedure whereby the diagnosis is arrived at by inference to the best possible explanation of the indicators. Consideration of alternative diagnoses is therefore a necessary part of justifying a diagnosis. Consideration of changes in conditions that may lead to changes in the diagnosis is a necessary part of an ongoing justification of a diagnosis. Both types of considerations are thus indicatory of the quality of the diagnostic procedure.

Dimension 4: Decisional Role

AI systems may play a host of roles in diagnostic procedures. They may serve the purpose of screening patients before the actual diagnostic procedures. As in Bob’s case, they may be part of the diagnostic procedures themselves. This involvement comes in degrees. They may be involved by providing limited input into the diagnostic process, but they may also play the role of deciding a diagnosis only to be ratified by a healthcare professional. Following diagnostic procedures, AI diagnostic systems may also be used as a “second opinion” [16].

The involvement of AI in health care may serve different interests including not only more accurate diagnostics but also more efficient (time) and more cost-efficient diagnostic procedures and so on. Any specific organizational implementation of an AI system in health care will be a balancing of such purposes and interests. Although a patient may share some – perhaps all – of these interests, the patient may not necessarily share the balancing of these interests in an actual organizational implementation of AI. Bob may be rather narrowly focused on getting the most accurate diagnosis, whereas the clinic may pursue the most accurate and cost-efficient procedure. On the assumption that in the present state of affairs increased diagnostic accuracy can only be achieved by combining the efforts of AI system and healthcare professionals, Bob may come to be diagnosed by an AI system embedded in organizational setup that will provide less accuracy than (1) what he could have achieved in an alternative organizational setup with AI and healthcare professionals, perhaps even less than (2) what he would have achieved under normal circumstances without the use of an AI diagnostic system.

A suboptimal organization of AI-supported diagnostics may obtain in the absence of conflicting interests. If increased diagnostic accuracy requires fruitful collaboration of man and machine, then it is of cardinal importance that healthcare professionals retain their ability to make their own independent diagnostic judgments and decisions. Yet, there is evidence indicating that clinical decision support systems may lead to the deskilling of healthcare professionals and automation bias, i.e., overreliance on the performance of such decision support systems [32–38]. If healthcare professionals likely substitute their own judgments with those of an AI system, then Bob and other patients may end up being de facto diagnosed solely by an AI system regardless of the organizational setup. However, there is also evidence suggesting that decision support systems may improve practitioner performance [39, 40]. Interestingly, one possible explanation of these differences in findings is that they reflect organizational differences between the contexts in which the influence of AI decision support on the

performance of healthcare professionals is being studied. Thus, these studies may ultimately be evidence in support of the more fundamental claim that the organization and the division of diagnostic labor do affect the quality of the diagnostic procedure.

The potential variations in diagnostic quality due to variations in organization raise questions related to the patients' right to distributive fairness and their right not to be harmed and the corresponding right to protect these rights. A setup, which in terms of diagnostic accuracy may leave patients worse off than what they could have been or than what they would under normal circumstances be (prior to the implementation of the AI system), is a setup that may be challenged on grounds of its distributive fairness. A setup that may affect diagnostic accuracy – and thus be the cause of over- and under-diagnoses and eventually over- and undertreatment – is a setup that may be challenged on grounds of its potential harm. We believe that patients should be able to defend themselves against distributive unfairness and harm by contesting the division of diagnostic labor between healthcare professionals and AI system in a specific organizational setup. In order to do so effectively, patients should have access to information about the (1) the role of the AI system in the diagnostic process, (2) the role of healthcare professionals in the diagnostic process, and not least (3) information about the objective/legal responsibility for the diagnostic procedures.

Further Issues and Aspects of the Right to Contest

By taking the right to contest as the fundamental reference point for the transparency requirements of AI diagnostics, we are taking a patient-centered approach. It is the right of the patient being subjected to concrete AI decision-making in a particular clinical setting to contest that determines the transparency requirements. As outlined in the introduction, there are alternative frameworks grounding transparency requirements. Thus, one may take a democratic approach grounding transparency requirements in the need of the wider public to be able to have informed discussions

about and hold decision makers accountable for the distributive principles embedded in AI systems and the power exercised over citizens by the use of AI. We believe, however, that the patient-centeredness of the right to contest is primary for two closely linked reasons: firstly, because democracy so to speak starts with an individual's right to contest decisions that may harm them or violate their rights. The right to contest is part of a set of rights establishing the ideal of a sovereign and autonomous individual that is "later" at a societal/communal level protected through the institutions of democracy and particular democratic rights. Bob's rights as citizen in a modern democracy starts by his right to protect himself whenever he is subjected to decisions that may violate his rights. Secondly, because the transparency requirements entailed by the right to contest seem to constitute the core of transparency requirements grounded in the protection of democracy. Public deliberations of the distribution principles of AI systems and the power they exercise over citizens would arguably exactly require insights into the personal and sensitive data used by such systems, the potential biases, the performance, and the organization and division of diagnostic labor between AI system and healthcare professionals. This is obviously less surprising if we are right that democracy starts by protecting the sovereignty and autonomy of the individual.

The scientific approach to transparency would in its purest form focus on algorithmic transparency and perhaps extend that to humanly understandable explainability. It is not only researchers or healthcare professionals who can find this approach valuable, but patients may also have an interest in this deeper understanding of the inner workings of the black box. However, focusing on algorithmic transparency and explainability is by definition a focus on the AI system in isolation and thereby excludes the many contextual factors that come before, during, and after the use of the AI system. It also potentially obscures those instances of bias that are not easily identifiable in the algorithm itself.

While a patient-centered and value-based right to contest AI decisions in health care may be appropriate, there is certainly need of further

work. Two issues are particularly relevant for further work and discussion. On the one hand, the question of the completeness of the information requirements laid out in this chapter. We have tried to show the relevance of particular types of information for Bob's right to contest. These conceptual links may be disputed. Moreover, there are other types of information that is vital and more pertinent for Bob's ability to contest the AI diagnostics in particular circumstances. On the other hand, the practical implications of the right to information about AI involvement in the diagnostic process. We have argued for the right to access to information, but who is responsible for retrieving and making this information available to patients, and how should it be made available? If the right to contest should have any practical relevance for patients, the information must be made available in ways taking into account their ability to and interests in processing such information and by extension the ability of healthcare professionals to process and communicate the information. However, we contend that the regulatory power of a right to contest – the influence of this right on the development and use of AI in health care – does not come solely from patients exercising this power, but from the very existence of this right, i.e., from the possibility that patients may request information about an AI diagnostic system and on this ground contest its decisions. We do not doubt, however, that patients in a number of cases will be motivated to contest diagnostic decisions or treatment choices. Bob may be one such patient.

References

1. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health.* 2019;1(6):e271–97.
2. Lipton ZC. The mythos of model interpretability. *ArXiv160603490 Cs Stat* [Internet]. 2016 Jun 10 [cited 2019 May 22].
3. Burrell J. How the machine ‘thinks’: understanding opacity in machine learning algorithms. *Big Data Soc.* 2016;3(1):2053951715622512.
4. Doshi-Velez F, Kim B. Considerations for evaluation and generalization in interpretable machine learning. In: Escalante HJ, Escalera S, Guyon I, Baró X, Güclü Y, Güclü U, et al., editors. Explainable and interpretable models in computer vision and machine learning [Internet]. Cham: Springer International Publishing; 2018 [cited 2019 May 22]. p. 3–17.
5. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access.* 2018;6:52138–60.
6. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining – KDD ’15 [Internet]. Sydney: ACM Press; 2015 [cited 2019 May 22]. p. 1721–30.
7. London AJ. Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hast Cent Rep.* 2019;49(1):15–21.
8. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell.* 2019;1(9):389–99.
9. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance) [Internet]. OJ L, 32016R0679 May 4, 2016.
10. Goodman B, Flaxman S. European Union regulations on algorithmic decision making and a “right to explanation.” *AI Mag* 2017;38(3):50–57.
11. Edwards L, Veale M. Enslaving the algorithm: from a “right to an explanation” to a “right to better decisions”? *IEEE Secur Priv.* 2018;16(3):46–54.
12. Wachter S, Mittelstadt B, Floridi L. Why a right to explanation of automated decision-making does not exist in the general data protection regulation [Internet]. Rochester: Social Science Research Network; 2016 [cited 2017 Apr 12]. Report No.: ID 2903469.
13. Richens JG, Lee CM, Johri S. Improving the accuracy of medical diagnosis with causal machine learning. *Nat Commun.* 2020;11(1):3923.
14. Daniels N, Sabin J. Limits to health care: fair procedures, democratic deliberation, and the legitimacy problem for insurers. *Philos Public Aff.* 1997;26(4): 303–50.
15. Kahneman D, Tversky A. Choices, values, and frames. 1st ed. Cambridge University Press; 2000. 860 p.
16. Ploug T, Holm S. The four dimensions of contestable AI diagnostics – a patient-centric approach to explainable AI. *Artif Intell Med.* 2020;107:101901.
17. Ploug T. In Defence of informed consent for health record research – why arguments from ‘easy rescue’, ‘no harm’ and ‘consent bias’ fail. *BMC Med Ethics.* 2020;21(1):75.
18. Shen N, Bernier T, Sequeira L, Strauss J, Silver MP, Carter-Langford A, et al. Understanding the patient privacy perspective on health information exchange: a systematic review. *Int J Med Inform.* 2019;125:1–12.
19. Esmaeilzadeh P, Sambasivan M. Patients’ support for health information exchange: a literature review and

- classification of key factors. *BMC Med Inform Decis Mak.* 2017;17(1):33.
20. Sankar P, Mora S, Merz JF, Jones NL. Patient perspectives of medical confidentiality: a review of the literature. *J Gen Intern Med.* 2003;18(8):659–69.
21. Scott IA. Hope, hype and harms of Big Data. *Intern Med J.* 2019;49(1):126–9.
22. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 2019;366 (6464):447–53.
23. Bobrowski D, Joshi H. Unmasking A.I.’s bias in healthcare: the need for diverse data. *Univ Tor Med J.* 2019;96(1):48–50.
24. Cabitz F, Ciucci D, Rasoini R. A giant with feet of clay: on the validity of the data that feed machine learning in medicine. *ArXiv170606838 Cs Stat [Internet].* 2018 May 14 [cited 2019 Nov 17].
25. Char DS, Shah NH, Magnus D. Implementing machine learning in health care – addressing ethical challenges. *N Engl J Med.* 2018;378(11):981–3.
26. Chen JH, Asch SM. Machine learning and prediction in medicine – beyond the peak of inflated expectations. *N Engl J Med.* 2017;376(26):2507–9.
27. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med.* 2018;178(11):1544–7.
28. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med [Internet].* 2019 [cited 2019 Nov 12].
29. Mac Namee B, Cunningham P, Byrne S, Corrigan OI. The problem of bias in training data in regression problems in medical decision support. *Artif Intell Med.* 2002;24(1):51–70.
30. Lichter AS. Conflict of interest and the integrity of the medical profession. *JAMA.* 2017;317(17):1725.
31. Mitchell AP, Trivedi NU, Gennarelli RL, Chimonas S, Tabatabai SM, Goldberg J, et al. Are financial payments from the pharmaceutical industry associated with physician prescribing? *Ann Intern Med [Internet].* 2020 [cited 2021 Feb 3].
32. Cabitz F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *JAMA.* 2017;318(6):517–8.
33. Tsai TL, Fridsma DB, Gatti G. Computer decision support as a source of interpretation error: the case of electrocardiograms. *J Am Med Inform Assoc.* 2003;10 (5):478–83.
34. Povyakalo AA, Alberdi E, Strigini L, Ayton P. How to discriminate between computer-aided and computer-hindered decisions: a case study in mammography. *Med Decis Mak.* 2013;33(1):98–107.
35. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J Am Med Inform Assoc.* 2012;19(1):121–7.
36. Goddard K, Roudsari A, Wyatt JC. Automation bias: empirical results assessing influencing factors. *Int J Med Inform.* 2014;83(5):368–75.
37. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. *J Am Med Inform Assoc.* 2017;24(2):423–31.
38. Zerilli J, Knott A, Maclaurin J, Gavaghan C. Transparency in algorithmic and human decision-making: is there a double standard? *Philos Technol.* 2019;32(4):661–83.
39. Garg AX, Adhikari NKJ, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *JAMA.* 2005;293(10):1223–38.
40. Sullivan F, Wyatt JC. How decision support tools help define clinical problems. *BMJ.* 2005;331 (7520):831–3.



Pierangela Bruno, Francesco Calimeri, and Gianluigi Greco

Contents

Introduction	240
Clinical Data	240
Electronic Health Record	241
Omics Data	242
Diagnostic-Related Information and Treatment Information	243
Data Processing	244
Missing Value Imputation	244
Dimensionality Reduction	244
Different Omics Processing	245
Existing and Emerging Applications in Medical Diagnosis	246
Clinical Data Application	246
Omics Data Application	247
Explainability of Results	248
Future Perspectives	249
Conclusion	250
References	250

Abstract

Providing accurate diagnoses of diseases and maximizing the effectiveness of treatments requires, in general, complex analyses of many clinical, omics, and pathological data. Making a fruitful use of such data is not straightforward, as they are usually stored in electronic health records (EHRs) that need to be properly handled and processed in order to successfully perform medical diagnosis. In recent years, machine learning and deep learning techniques have emerged as powerful tools

Supplementary Information: The online version of this chapter (https://doi.org/10.1007/978-3-030-64573-1_32) contains supplementary material, which is available to authorized users.

P. Bruno (✉) · F. Calimeri · G. Greco
Department of Mathematics and Computer Science,
University of Calabria, Rende, Italy
e-mail: bruno@mat.unical.it; francesco.calimeri@unical.it;
calimeri@mat.unical.it; gianluigi.greco@unical.it

to perform specific disease detection and classification using EHRs data, thus providing significant clinical decision support. However, approaches based on such techniques suffer from the lack of proper means for interpreting the choices made by the models, especially in the case of deep-learning ones. In this chapter, we describe clinical and omics data along with the popular processing operations performed to improve the medical analyses. We present the most common used algorithms in automatic medical diagnosis and the advance in explainability of machine learning-based systems to validate healthcare decision-making.

Keywords

Medical diagnosis · Health care · Clinical data · Electronic health records · Omics data · Machine learning · Deep learning · Explainability

Introduction

A large amount of patient-related information is collected by healthcare operators in their everyday activities, which span over a wide spectrum of medical processes, such as wellness check-ups or examinations at healthcare hospitals or medical offices, just to name a few. For instance, when a patient undergoes a medical examination for the first time, the physician usually creates a patient file including his medical history, current treatments, medications, diagnosis, and other relevant information [1]. Considering that disease diagnosis is crucial for health condition monitoring, it is natural to envisage that such large amount of data can be profitably used to guide data-driven disease classification tasks in the quest for early and accurate diagnoses, taking care of the complex interactions among clinical, biological, and pathological variables. Indeed, with the aim of identifying the best services and treatments for the patients, recent advances in medicine have proposed various models for personalized, predictive, and preventive medicine that make use of electronic health records (EHRs) and high-dimensional omics data [2]. In particular,

the use of EHRs facilitates healthcare delivery, by possibly contributing to improving healthcare quality with more accurate decisions. However, accessing and using EHRs and omics data can be rather challenging in practice, because they are heterogeneous and usually stored in different data formats. Therefore, these data need to be collected, cleaned, and stored in a proper way. On top of the consolidated data, relevant and meaningful knowledge for patient care can be then extracted by means of advanced Artificial Intelligence (AI) methods and Machine Learning (ML) approaches.

In this chapter, we focus on the definition of precision medicine and precision health based on clinical and omic data, by considering in particular disease prediction and prevention. We present a survey of recent and novel works in medical diagnosis, along with a detailed discussion about the relevance of the explainability of the results to validate the treatment. In particular, we firstly focus on medical applications based on EHRs. Then, we present recent advances in omics technologies (i.e., genomics, transcriptomics, epigenomic, proteomics, and metabolomics) to enable personalized medicine [83].

The remainder of the chapter is structured as follows. A detailed description of the salient features of clinical data is provided in section “[Clinical Data](#).” Different techniques used in data processing are detailed in section “[Data Processing](#).” Then, emerging applications in Medical Diagnosis are discussed in section “[Existing and Emerging Applications in Medical Diagnosis](#),” while the problem of providing explainable results is detailed in section “[Explainability of Results](#).” Eventually, a number of future perspectives are sketched in section “[Future Perspectives](#).”

Clinical Data

Health-related information usually refers to clinical data, and it is associated with regular patient care or it comes as a part of a clinical trial program. Clinical data represent an important resource to enable and guide the acquisition of novel knowledge (for instance, to provide outcome and therapies predictions) and best practices

(for instance, to ensure data completeness, reliability, and correctness) in healthcare [3]. Clinical data are also important to seek timely treatments and appropriate care to the patient, and to enable a (sort of) self-learning system that is able to continuously improve quality of care. Indeed, by considering the importance of health condition monitoring for early and accurate medical diagnosis, learning from practice is crucial. Clinical data are collected and translated into EHRs; these include administrative and demographic information, diagnoses, treatments, prescription drugs, physiologic monitoring data, and other health information, as reported in section “[Electronic Health Record](#).” The use of EHRs in clinical research studies provides several advantages, such as defining the most suitable treatments, reducing hospitalization cost, and providing personalized medicine. In particular, EHRs are used for various data science and machine learning studies, including statistical analysis of diseases, risk prediction for breast cancers [4], classification of heart disease and diabetes [5], and diagnosis of rare pathologies [6].

In the last decade, different studies have shown that it is possible to discover relevant information and distinctive attributes related to specific diseases by properly analyzing and combining omics and EHRs data [7–10]. Omics studies are used to analyze types of molecules in samples which can be measured in terms of character and quantity as a whole, with the aim of investigating the patterns or relations to the sample attributes [11]. Thanks to the comprehensiveness of these data, it is possible to generate or confirm hypotheses on biological or medical conditions [12] and to provide a basis for precision medicine [13].

Precision health, an evolution of precision medicine, combines omics data with lifestyle, clinical data, and environmental factors. These data can be used to identify and predict disease diagnosis, treatment, prevention, and individualized early diagnosis [14]. In particular, a personalization of medical treatment based on specific patient characteristics is possible by improving the understanding of the physiological and biological mechanisms of disease, responsible for the spreading of omics data, and by developing patient-based algorithms [15]. Thanks to these

advances, the work on precision health is growing quickly, thereby allowing an improvement of outcomes and reducing unnecessary treatment. However, clinical data as well as omics data present several issues that can have a substantial impact on study results, especially in terms of quality, such as incompleteness (e.g., missing information), inconsistency (e.g., information mismatch between various or within the same data source), and inaccuracy (e.g., nonspecific, non-standards-based, inexact, incorrect, or imprecise information) [16, 90].

In this context, Omics Data Management (ODM) and Clinical Data Management (CDM) play a crucial role in generating high-quality, reliable, and statistically sound data from clinical trials [17]. On the one hand, considering that omics data usage provides a prominent opportunity for diagnosis, or predictions of future diagnoses, ODM is used to address the uncertain or unexpected findings, many of which may have relevant impacts in medical diagnosis (e.g., the collection and processing of genome data that are not sufficiently standardized or valid); on the other hand, the main goal of CDM processes is to provide high-quality data by reducing the number of errors and missing data to improve data analysis [17].

In general, data management is the process of collecting, cleaning, and processing data according to standard rules. With this aim, the use of specific software applications and best practices can lead to data completeness and reliability. Sections “[Electronic Health Record](#)” and “[Omics Data](#)” report a detailed description of electronic health record and omics data, respectively, by highlighting the main characteristics and issues.

Electronic Health Record

Electronic health records (EHR) build the digital version of a person’s medical information and history, which is maintained over time by a healthcare provider. EHRs are used to timely and consistently collect patient information to provide more extensive and accurate clinical care, and in predicting future outcomes based on both individual-related and population-related data. It has been shown that a proper use of EHRs can

improve healthcare quality, with benefits including secure long-term storage, consistency, standardization, and accessibility of patient information [18]. EHRs can include structured and unstructured data such as demographics, medical and surgical histories, diagnoses and procedures, patient examinations, and results from various clinical studies, as listed in Table 1. Structured data use a uniform format in the EHRs system with a controlled vocabulary and pre-determined values that make these data consistent and easily extractable. To the contrary, unstructured data do not use a standard format; indeed, healthcare providers can include free text regarding, for example, additional health information and details on clinical examinations, along with comments by the operators. Then, the same clinical information can be recorded in EHRs in different ways depending on the user, hence posing important challenges to computer analysis because of (1) heterogeneous data formats, (2) abundant typing and spelling errors, (3) violation of natural language grammar, and (4) rich domain-specific abbreviations, acronyms, and idiosyncrasies [13]. The unstructured data requires the use of additional tools, such as natural language processing (NLP), to standardize, codify, and extract relevant information [23].

Omics Data

Understanding of human health and diseases requires a proper interpretation of molecular interactions and variations at multiple levels such as genome, epigenome, transcriptome, proteome, and metabolome [24], which together go under

the name of omics data. In other words, omics refers to the collective technologies used to explore the roles, relationships, and actions of the various types of molecules that compose cells of an organism [25] and, potentially, are responsible for specific disease or condition [26]. The omics data and the description of each molecular profile are listed in Table 2.

The use of high-throughput data acquisition such as next-generation sequencing (NGS) and mass spectrometry (MS) allow to perform fast accumulation of omics data. These data can be used to identify medical biomarkers by cleaning raw data generated by NGS or MS, extracting molecular profiles, identifying statistically significant molecules, or defining models able to explain molecular interactions in a specific context. Specifically, genomics approaches have been used to identify the genes and genetic loci involved in the development of human diseases [27], epigenomics to find epigenetic markers [28], proteomics for proteins and peptides [29], and metabolomics for low-abundance metabolites [30]. These data, which are composed of a huge amount of data points, require standard data formats and publication guidelines [31] to improve data sharing, acquisition, analysis, and usage. The use of standard formats ensures unambiguous communications, clear experimental designs, treatments, and analyses with the aim to support the conclusions, guarantee an independent reproduction [31] and facilitate the creation of standardized public data repositories. The current state-of-the-art counts different methods to facilitate better exchange and integration of data such as the Minimum Information About a Microarray Experiment (MIAME) [32] to define a guideline

Table 1 Example of structured and unstructured data available as EHRs at patient-level

EHR data type	Category of data	Example of data
Structured	Demographics	Age, name, gender, contact information
	Allergies	Drug allergies and adverse reactions
	Vital signs	Height, weight, temperature, heart rate, pressure
	Diagnosis codes	International Classification of Diseases (ICD) used as diagnostic tool for diseases classification, health status monitoring and clinical purposes
Unstructured	Clinical notes	Progress notes, hospital admission notes, physical therapy, history and physical examination

Table 2 Some domains of omics standards are listed. The domain indicates the type of experimental data. Genomic measure DNA molecules, epigenomic, transcriptomic,

proteomic, and metabolomic studies the chemical states of DNA and its binding proteins, RNA, proteins, and metabolites, respectively [11]

Domain	Description
Genomic	The genome is the complete sequence of DNA in a cell or organism. Genomic is a technique used to improve diagnosis through identification of genomic conditions, which could improve clinical management, prevent complications, and promote health [7]
Transcriptomic	The transcriptome is the complete set of RNA transcripts from DNA in a cell or tissue. Transcriptomic is a technique used to measure the chemical states of DNA and its binding proteins, RNA, proteins, and metabolites, respectively [11] and study an organism's transcriptome, the sum of all of its RNA transcripts. The information content of an organism is recorded in the DNA of its genome and expressed through transcription, which is also used in disease diagnosis and profiling [19]
Epigenomic	The epigenome consists of reversible chemical modifications to the DNA, or to the histones that bind DNA. Epigenomic is a technique used to study the complete set of epigenetic modifications on the genetic material of a cell, known as the epigenome. Epigenetic patterns are potential useful biomarkers to detect cancer cells and to classify different disease types [20]
Proteomic	The proteome is the complete set of proteins expressed by a cell, tissue, or organism. Proteomic is a technique used to study proteins, which are vital parts of living organisms, with many functions. Proteomic methodologies enable the detection and quantitation of protein profiles associated with the disease state [21]
Metabolomic	The metabolome is the complete set of small molecule metabolites found within a biological sample. Metabolic is a technique used to study and monitor the presence and concentration of specific metabolites associated with a particular disease. Metabolomics methodologies provide the opportunity for the identification of clinically relevant biomarkers as it best mirrors the human phenotype [22]

for the minimum information required to describe a DNA microarray-based experiment, the Minimum Information About a Proteomics Experiment (MIAPE) [33] to define guidelines for proteomics studies, and the Minimum Information about a high throughput SEQuencing Experiment (MINSEQE) [34] to define guidelines for sequencing study.

Diagnostic-Related Information and Treatment Information

Diagnostic-related information can benefit patients in receiving the right treatment, helping healthcare operators to provide the most appropriate preventive interventions and therapeutic strategies [35]. Diagnostic information can help to avoid or shorten hospitalization, reducing inappropriate use of drugs and, consequently the economic cost, bringing to more efficient use of resources [23]. Although diagnostic information can help clinicians to identify certain biomarkers in the body, they do not allow any identification of insight into the interaction among information and the environment responsible for the disease.

The vast amount of digital data captured in EHRs in combination with the emergence of omics data offers novel research perspectives and opportunities in health systems to improve health management and enable the study of personalized medicine and treatment information. In this scenario, the use of omics data becomes necessary to transform the understanding and treatment of many diseases, helping doctors to customize therapies to the molecular profiles of individual patients. These data represent a crucial element in predicting biomarkers, in determining predisposition, diagnostic, and prognostic and identifying modifiable risk factors with the main objective to improve the traditional symptom-driven practice of medicine [36].

However, integrating clinical and omics data is a challenging task due to the difficulty in knowledge extraction from large and complex datasets. Such issues can be tackled through machine learning algorithms that can be adapted to specific settings and omic types [37]; indeed, a high number of researches have been carried out to examine the application of machine learning and deep learning techniques in the area of diagnostic-related and treatment information; some existing and emerging applications are described in section

“Existing and Emerging Applications in Medical Diagnosis.”

Data Processing

Omics and EHRs data can be difficult to analyze and computationally expensive to process due to high-dimensionality. In particular, omics data are composed of many dimensions/features much larger than the number of samples available, while EHR data usually contains a large sample size of high-dimensional data, but each individual sample is sparsely populated [13]. Also, these data can suffer from information quality problems, making data not really suitable for clinical research. In order to tackle these issues, data quality assessment and dimensionality reduction are needed to remove irrelevant and redundant data and preserve the characteristics of the original data. An overview of the data processing operations is shown in Fig. 1.

In the following, we first describe techniques used to handle missing values in section “[Missing Value Imputation](#).” Then, we illustrate dimensionality reduction technique in section “[Dimensionality Reduction](#).” And, eventually, we present further pre-processing approaches that are used to deal with omics data in section “[Different Omics Processing](#).”

Missing Value Imputation

In order to maximize the performance of data analysis, the quality of each omic and EHR data

must be carefully assessed to ensure that measurements are reproducible. Since missing values can affect the quality of the results, several approaches were proposed to tackle this issue. One of the most common approaches is called single imputation, which relies on the removal of the missing rows, by ignoring subjects with incomplete information or replacing the missing items with plausible values (e.g., means of the observed cases). This procedure presents some limitations due to the high proportion of subjects discarded [38], the distortion of the distribution of the variables, and the introduction of additional biases [39]. More robust approaches were presented in the state-of-the-art such as multiple imputation [39], inverse Probability Weighting [40], expectation-maximization [41], Multivariate Imputation by Chained Equation [42].

Dimensionality Reduction

Feature extraction and selection methods are used to maximize performance analysis and improve result understandability. **Feature Selection** selects features in input that contains relevant information for solving the particular problem. Instead, **Feature Extraction** is used to transform the input space into a low-dimensional subspace that preserves the most relevant information [43].

Feature selection techniques consist of filtering, wrapper, or embedded methods [13]. In the **filter** approach, each feature is evaluated individually using its general statistical properties [44],

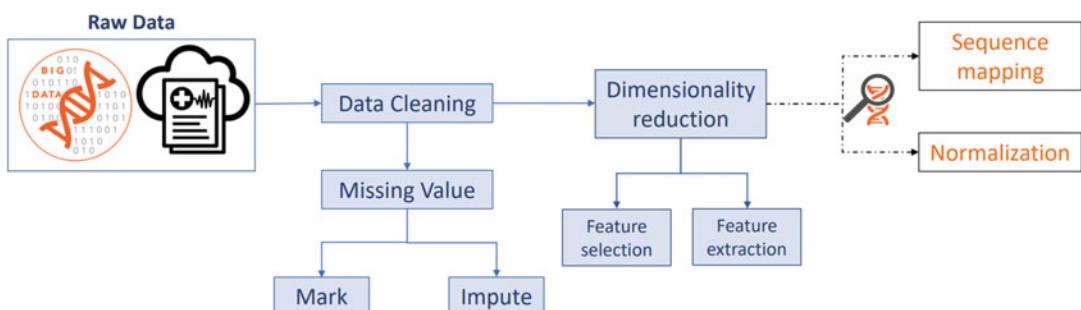


Fig. 1 Workflow of data processing steps. Data cleaning and dimensionality reduction are performed on raw data. Sequence mapping and normalization are usually used to handle the complexity of omics data

making the approach faster, without explicit class labeling [13]. Among several filter approaches, we selected and described the most common. Mutual Information (MI) measures the level of dependence between two random features (i.e., the amount of information that variable V1 knows about another variable V2) [45]. Information Gain (IG) evaluates the gain of each feature to a certain class, by computing the entropy (i.e., level of impurity), such that the feature with higher information is the much related, while the unrelated feature offers no information [46]. Minimum Redundancy Maximum Relevance (mRMR) [44] iteratively selects features with the maximum relevance, decreasing the redundancies within each class, such that any mutually exclusive features are selected [47].

The **wrapper** approach uses learning techniques to select the optimal feature subset. This approach typically requires high computational costs and it is affected by the risk of overfitting, but it shows better performances than the filter approach [48]. Genetic Algorithm (GA), Ant Colony Optimization (ACO), and Particle Swarm Optimization (PSO) are examples of wrapper approaches inspired by natural evolution, ant colonies and flocks of birds, respectively, to generate a population of features that optimize the solution.

The **embedded** method integrates machine learning algorithms with recursive feature elimination [13]. These methods are less computationally expensive and less prone to overfitting than the wrapper ones. An example of an embedded method is Support vector machines (SVM), which is a supervised machine learning algorithm used to search for a hyperplane that optimally divides the tuples from one class to another. Therefore, SVM allows us to identify the main features used in classification and remove the not important ones [49].

Among the various feature extraction techniques, principal component analysis (PCA) is a commonly used method. PCA is an orthogonal linear transformation that converts variables into a new smaller set, called principal components. The number of principal components is less than or equal to the number of original feature variables [50].

Different Omics Processing

Depending on the objective being defined and on the dataset itself, other preprocessing approaches may be appropriate to handle the complexity of omics data, such as (I) sequence mapping and (II) normalization.

(I) **Sequence mapping** is the process of comparing millions of sequences generated, for example, by NGS against the reference genome to obtain one alignment between each read and the genome. Mapping is fundamental in NGS and MS analysis, representing the basis for further analysis, e.g., to estimate the abundance of transcripts and variant detection [51]. The common NGS mapping tools are based on a hash table or index-based algorithms, while heuristic-based aligners are demonstrated to be less expensive in terms of computation [51], such as Genomic Mapping and Alignment Program (GMAP) [52] and Burrows-Wheeler Alignment (BWA) [53]. In MS data, peak alignment can correct drifts to ensure accurate mass across samples. Indeed, without alignment, the same peak (e.g., the same peptide) can have different values of mass-to-charge ratio (m/z) across samples [84].

(II) **Normalization** plays an important role in the omics data preprocessing to remove unwanted systematic bias while maintaining real biological differences in the observed datasets. In this context, different methods have been used based on several statistical models (e.g., unit norm, median, and quantile), scaling methods (e.g., auto-scaling, range scaling, Pareto scaling, vast scaling, and level scaling), and data transformation (e.g., log and power) [54]. Other approaches are used to normalize the samples by adjusting their variability. For instance, Locally Weighted Scatterplot Smoothing (LOWESS) algorithm [55] is used to remove the bias present in the data which frequently shows a deviation from zero for low-intensity spots. Normalization also enables the comparison of different samples in MS, by considering that the absolute peak values of different fractions of the spectrum could be incomparable. Normalizing spectrum allows us to identify and remove sources of systematic variation in MS data which can depend, for example, on varying amounts of sample or variation in the instrument detector sensitivity.

Existing and Emerging Applications in Medical Diagnosis

Recent improvements in machine learning (ML) and deep learning (DL) approaches provided useful techniques to recognize patterns from clinical and omics data and to predict future outcomes of the patients. A conceptual model to illustrate the diagnostic process is shown in Fig. 2. In fact, several works have already shown the validity of these methods to provide improved and more generalizable risk prediction models: by taking advantage of high-dimensionality data, such as EHRs or omics data, these algorithms can effectively determine the combinations of variables that are able to predict a reliable outcome.

In this section, we will describe the latest advances in ML and DL on EHRs and omics data, with the latter being considered from the molecular profiles perspectives: genomics, transcriptomics, epigenomics, proteomics, and metabolomics.

Clinical Data Application

Several works have been proposed to perform diagnosis prediction using clinical data collected during a single medical check-up. Corey et al. [56]

defined different machine learning algorithms including penalized logistic regression, random forest models, and extreme gradient boosted decision trees to identify high-risk surgical patients from EHRs [85]. The experimental analysis showed that the best result was produced by penalized logistic regression models with an AUC value of 0.92. Instead, a DL-based approach was proposed by Miotto et al. [57]. They used a deep neural network composed of a stack of denoising autoencoders to process EHRs in an unsupervised manner that captured stable structures and regular patterns in a variety of clinical risk-prediction tasks, including, for example, diabetes mellitus with complications, cancer of rectum and anus, cancer of liver and intrahepatic bile duct, congestive heart failure (non-hypertensive). The proposed approach outperformed state-of-the-art methods based on raw EHR data (i.e., no feature learning applied to EHR data). A limitation of these approaches is, however, that they did not rely on temporal information to improve the performance of predictive models, making them not applicable in real clinical settings.

To tackle time dependencies in EHRs, Rohit et al. [58] presented a continual prediction framework based on logistic regression to predict acute kidney injury (AKI) before the development of disease at any time during the hospitalization. In

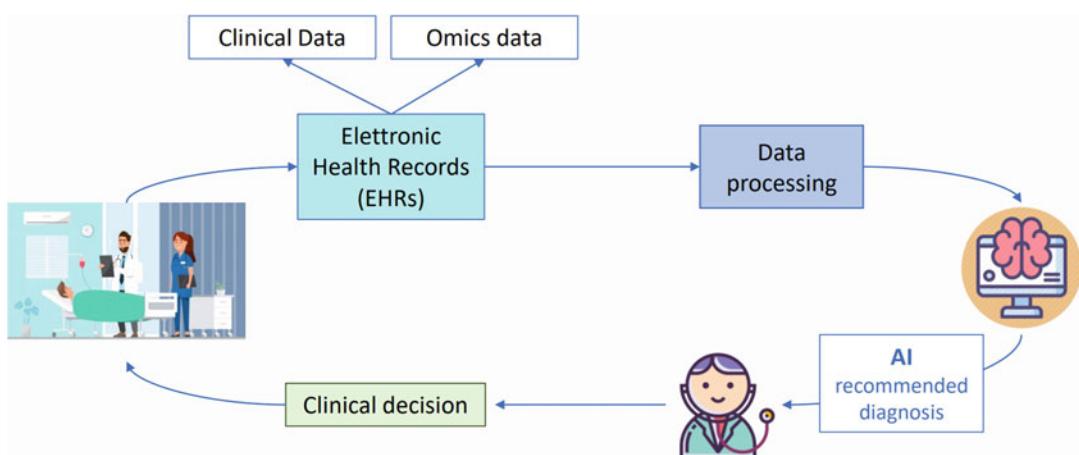


Fig. 2 EHRs are composed of clinical data and omics data that are collected during medical examinations. Data processing operations are performed to prepare data for AI

techniques. The output is used by healthcare providers to improve the diagnosis and provide a clinical decision

particular, the proposed model continually predicts over the entire hospital stay whenever any patient variable changes (e.g. AKI occurrence). A limitation of the work is related to the data used to predict the disease; indeed, the model was obtained by using only the structured part of EHR, discarding the unstructured component. Mehak et al. [59] proposed a recurrent neural network (RNN) architecture with long short-term memory (LSTM), which learns the patient representation from the temporal data collected over various visits of the patient, to predict future obesity patterns from childrens' medical history. Jeong et al. [60] proposed a similar approach called multi-scale temporal memory (MTM) with the aim of modeling clinical events at different time scales in EHRs. Thanks to MTM, information about past events on different time-scales are collected and used to perform on-the-fly prediction. The authors showed that the approach, combined with different patient states, can cover different temporal aspects of patient states, achieving an accuracy improvement of 4.6% than baseline approaches and of 16% the prediction based on LSTM approach. A limitation of these approaches is related to the limited size of the observation windows in the EHRs data. Finally, Che et al. [61] develop a variation of the recurrent GRU cell (GRU-D) which handles missing values in clinical time series by incorporating time intervals directly inside the GRU architecture. Indeed, missing values are considered as one of the principal reasons for the loss of precision. Their GRU-D networks showed an improvement of AUC on two real-world healthcare datasets for classification and mortality prediction tasks.

Omics Data Application

Genomic data: DL technologies can be used to predict gene expression and to study interaction between genomes and diseases. Chen et al. [62] exploited a DL method to infer the expression of target genes from the expression of landmark genes. They used the microarray-based Gene Expression Omnibus dataset to train the model, achieving a performance that is significantly

better than logistic regression, with an improvement of 15.33% on mean absolute error. Chen et al. [63] developed a DL-based approach to explore genomes and disease relationships. They performed jointly supervised classification, unsupervised clustering, and dimensionality reduction on high-dimensional genomic data to identify cancer subtypes. This approach was tested on breast cancer and outperformed the existing methods, identifying more robust subtypes using fewer genes. A limitation of the work is related to generalization capability. The performance decreases in predicting other types of cancers where molecular subtypes have not yet been well established. Bruno et al. [64] presented a framework for automated diagnosis based on high-dimensional gene expression. They performed data reduction to transform high-dimensional representations of data into a lower-dimensional space. Eventually, they relied on different data visualization techniques to convert complex pieces of information into 2-D images, which are in turn used to perform DL-based diagnosis. They also tested the framework on clinical data to perform diagnosis. The approach obtained a good performance, showing a prediction recall value between 0.91 and 0.99. However, performances decrease in predicting patients that share similar variations in gene expression or clinical information.

Transcriptomic data: DL technologies can be also used to identify the association between RNA and disease. Thomas et al. [65] proposed a deep belief neural network (deepBN) to perform the classification of pre-miRNAs. DeepBN is composed of an unsupervised stage with hidden layers pre-trained as restricted Boltzmann machines (RBMs), followed by a supervised tuning of the network [86]. The approach achieved a level of accuracy of 0.97. Bobak et al. [66] presented a data analysis framework that directly integrates multiple publicly-available expression array datasets to identify a gene signature for the diagnosis of tuberculosis. The authors evaluated different machine learning algorithms including random forest, support vector machine with the polynomial kernel, and partial least square discriminant analysis. According to the analyses,

the authors proved that the best result was obtained using random forest with an accuracy value of 0.95.

Epigenomic data: In the context of predict epigenetic effects of DNA sequence alterations (e.g., chromatin accessibility, DNA methylation and histone modifications), DL methods can be used too [87]. Quang et al. [67] proposed a hybrid approach based on Convolutional neural network (CNN) and LSTM to predict the chromatin effects of noncoding DNA sequence alterations [87]. This approach allows to simultaneously learn motifs and a complex regulatory grammar between the motifs. The authors showed that the approach outperforms other methods for predicting the properties and function of DNA sequences across several metrics, including AUC. A limitation of the approach is related to the dimension of sequence data used to train the network; indeed, it can only process sequences of constant length with static output. Yin et al. [68] proposed a DL approach for integrating sequence information and chromatin data with the aim to perform prediction of modification sites specific to different histone markers. The architecture is composed of three modules corresponding to a DNA sequence, chromosome accessibility, and a joint module, respectively. The approach outperformed several baseline methods in a series of comprehensive validation experiments [68]. The authors did not incorporate recurrent neural network architecture, such as long short-term memory units, that, considering the sequential natural DNA fragments, may improve the performance.

Proteomic data: DL technologies are also effective in the identification of protein structures and protein contact map prediction. Wang et al. [69] presented a new DL method to improve protein contact prediction by integrating both evolutionary coupling (EC) and an ultra-deep neural network to preserve sequence information. The approach achieved the highest F1 score on protein structure prediction (CASP). The prediction accuracy could be even better if the authors train a model with more layers. Liang et al. [70] proposed DL algorithms with stacked autoencoders for the analysis of FLT3-ITD mutation in acute leukemia patients. The authors relied on a dimensionality

reduction algorithm to reduce the number of proteins. The approach achieved an accuracy of 0.97.

Metabolomic data: Finally, DL technologies can be used to capture the metabolic features of complex traits and predict metabolic pathways. Stamate et al. [71] relied on several state-of-the-art algorithms, such as DL, Extreme Gradient Boosting (XGBoost), and Random Forest (RF), to obtain high accuracy models to predict Alzheimer's disease (AD) versus cognitively normal (CN) with metabolites as predictors. The implemented framework captured metabolic complexity in Alzheimer's disease (AD), achieving an AUC value of 0.88 and 0.85 using the XGBoost and DL-based approach, respectively. A limitation of the study is that the authors did not include an external validation due to the size of the cohort. Baranwal et al. [72] proposed the use of the RF classifier for metabolic pathway prediction. The input of the classifier was extracted from molecular structures in SMILES format using graph convolutional networks (GCN). This approach achieved 95% accuracy in predicting metabolic pathways. The authors showed that the approach is able to estimate the relative contribution of metabolites in distinguishing pathway classes.

Explainability of Results

Managing clinical and public health decision support systems requires to face several challenges. One of the main challenges is the lack of sufficient explanation of diagnostic and therapeutic solutions and interventions. Indeed, ML and DL-based methods can be more difficult to explain and justify to human users when compared to classic analytical models [73]. Indeed, although a great predictive ability, most of these techniques output complex information networks, which makes the decision process difficult to explain (i.e., the well-known “black box problem”) [74].

Recently, several approaches were proposed to understand the behavior of ML and DL-based systems, referred to as Explainable Artificial Intelligence (XAI). XAI is an important tool in medicine and health care to ensure that the results

obtained by AI methods are sound, correct, and justifiable in order to help healthcare providers in making better decisions [73]. Thus, providing an explainability in the process of making qualified decisions can improve and justify the treatment and health intervention, and help in translating inferred knowledge into particular hypotheses that can be tested with real-life experiments.

Among several approaches, the attention mechanism is one of the most widely used in Deep Learning research in the last decade, which is a method used to analyze the models capability and highlight the most relevant information used in the prediction task. Park et al. [75] proposed a novel disease prediction method named EHR History-based prediction using Attention Network (EHAN). The approach was based on the recurrent neural network (RNN) to predict vascular disease from EHR data. They included a gradient-weighted class activation mapping (Grad-CAM) [88] to visualize the class-specific attention-weights and to provide interpretable results by explicitly weighting significant features and visualizing them. Similarly, Bruno et al. [76] proposed the use of GradCAM and Guided-GradCAM to analyze the internal processes performed by a neural network during the task of automatic medical diagnosis based on images representing the high-dimensionality of gene expression and clinical data. Hu et al. [77] proposed an attention-based deep learning framework, named DeepHINT, to provide mechanistic explanations on accurate prediction of HIV performed on the genomic sequence data. This approach was used to reveal important sequence positions from prediction results and thus provide important insights about the observed genomic. Choi et al. [78] introduced a reverse time attention model (RETAIN), which uses two attention models to detect significant clinical variables within past visits (e.g., key diagnoses). RETAIN is able to preserve interpretability, mimic the behavior of healthcare providers, and incorporate sequential information.

Other approaches are focused on analyzing the level of contribution of the input features to the output prediction to provide interpretability to ML models. Shrikumar et al. [79] presented Deep

Learning Important FeaTures (DeepLIFT), which is a backpropagation-based interpretability approach. In particular, the approach calculates the output prediction of a network on a specific input by backpropagation algorithm to discover the importance of each feature. Lundberg et al. [80] proposed SHapley Additive exPlanations (SHAP) framework which uses a combination of feature contributions and game theory for breaking down the prediction with the aim to show the impact of each input feature. The overall rating of the feature contribution to the model is then achieved by aggregating the SHAP values over observations [89].

Future Perspectives

In order to improve healthcare assistance, clinicians need to make decisions with a high degree of certainty. AI models emerged in the literature as very promising approaches to improve the quality of prognostic and to support the whole diagnostic process. However, since correct clinical decision-making can influence the prognosis and affect treatments, healthcare providers need to be able to understand and trust the predictions and recommendations made by AI-based systems. But this is rather challenging in practice, since AI models pose even questions about their (human) interpretability. Investigating methods to clearly interpret (and then trust) the result of AI algorithms (especially, based on deep learning approaches) constitutes a very important avenue of further research.

Furthermore, although AI-based systems can provide promising results, many patients do not respond to drug treatment due to the complexity of diseases, which can be characterized by different and altered information among patients with the same diagnosis [81]. For this reason, the diagnosis, as well as the treatment plan, can be prone to error. In this context, the use of Digital Twins (DTs) shows promising results to improve the precision of the treatments, for it enables a virtual representation of the human body and with its organs where the effects of drugs and treatments can be studied. The use of DTs allows us to

integrate the information relevant to pathogenesis, with the aim of predicting the outcome of specific procedures by synthesizing and tracking patient data. Moreover, DTs can provide assistance in determining the proper treatment and the best drug among the thousands possible to treat a certain disease [82]. So, it is not hard to envisage that in the future, DTs and AI-based systems will help clinicians to properly and accurately optimize the performance of treatment plans based on specific patient characteristics.

Conclusion

This chapter reported a survey of the current status of research and literature related to AI-based applications in precision medicine and health via clinical and omic data. Thanks to these techniques – such as ML and DL it is possible to find insights into diagnostics, care processes, and treatment variability, identifying disease-related biomarkers and improving patient outcome.

The ongoing research shows very promising perspectives to enhance disease prediction and prevention and paves the way to digital representations of patients (i.e., the use of *Digital Twins* in healthcare) which could be essential for tasks like defining customized treatment decisions, predicting outcomes, identifying the best drugs among all possibilities for a certain disease, virtually testing several treatments (in advance), thus reducing impact on patients and risks. In future, Digital Twins and AI-based systems could significantly help clinicians to properly optimize the performance of treatment plans, based on specific patient characteristics.

References

- Evans JA. Electronic medical records system. Google Patents; 1999. US Patent 5,924,074.
- Ristevski B, Chen M. Big data analytics in medicine and healthcare. *J Integr Bioinform.* 2018;15(3):20170030.
- McGinnis JM, Olsen L, Goolsby WA, Grossmann C, et al. Clinical data as the basic staple of health learning: creating and protecting a public good: workshop summary. National Academies Press; 2011.
- Li R, Chen Y, Ritchie MD, Moore JH. Electronic health records and polygenic risk scores for predicting disease risk. *Nat Rev Genet.* 2020;21:493–502.
- Brisimi TS, Xu T, Wang T, Dai W, Adams WG, Paschalidis IC. Predicting chronic disease hospitalizations from electronic health records: an interpretable classification approach. *Proc IEEE.* 2018;106(4):690–707.
- Garclon N, Burgun A, Salomon R, Neuraz A. Electronic health records for the diagnosis of rare diseases. *Kidney Int.* 2020;97(4):676–86.
- Wise AL, Manolio TA, Mensah GA, Peterson JF, Roden DM, Tamburro C, et al. Genomic medicine for undiagnosed diseases. *Lancet.* 2019;394(10197):533–40.
- Bruno P, Calimeri F. Using heatmaps for deep learning based disease classification. In: 2019 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB). IEEE; 2019. p. 1–7.
- Zhu B, Song N, Shen R, Arora A, Machiela MJ, Song L, et al. Integrating clinical and multiple omics data for prognostic assessment across human cancers. *Sci Rep.* 2017;7(1):1–13.
- Oromendia A, Ismailgeci D, Cioffi M, Donnelly T, Bojmar L, Jyazbek J, et al. Error-free, automated data integration of exosome cargo protein data with extensive clinical data in an ongoing, multi-omic translational research study. *Proc Am Soc Clin Oncol.* 2020;38:e16743.
- Yamada R, Okada D, Wang J, Basak T, Koyama S. Interpretation of omics data analyses. *J Hum Genet.* 2020;66:93–102.
- Yu XT, Zeng T. Integrative analysis of omics big data. In: Computational systems biology. Springer; 2018. p. 109–35.
- Wu PY, Cheng CW, Kaddi CD, Venugopalan J, Hoffman R, Wang MD. Omic and electronic health record big data analytics for precision medicine. *IEEE Trans Biomed Eng.* 2016;64(2):263–73.
- Fu MR, Kurnat-Thoma E, Starkweather A, Henderson WA, Cashion AK, Williams JK, et al. Precision health: a nursing perspective. *Int J Nurs Sci.* 2020;7(1):5–12.
- Madhavan S, Subramaniam S, Brown TD, Chen JL. Art and challenges of precision medicine: interpreting and integrating genomic data into clinical practice. *Am Soc Clin Oncol Educ Book.* 2018;38:546–53.
- Ford E, Rooney P, Hurley P, Oliver S, Bremner S, Cassell J. Can the use of Bayesian analysis methods correct for incompleteness in electronic health records diagnosis data? Development of a novel method using simulated and real-life clinical data. *Front Public Health.* 2020;8:54.
- Krishnankutty B, Bellary S, Kumar NB, Moodahadu LS. Data management in clinical research: an overview. *Indian J Pharmacol.* 2012;44(2):168.
- Howe JL, Adams KT, Hettinger AZ, Ratwani RM. Electronic health record usability issues and potential contribution to patient harm. *JAMA.* 2018;319(12):1276–8.

19. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. *PLoS Comput Biol.* 2017;13(5):e1005457.
20. Weichenhan D, Lipka DB, Lutsik P, Goyal A, Plass C. Epigenomic technologies for precision oncology. In: *Seminars in cancer biology*. Elsevier; 2020.
21. Clark DJ, Zhang H. Proteomic approaches for characterizing renal cell carcinoma. *Clin Proteomics.* 2020;17(1):1–18.
22. Njoku K, Sutton CJ, Whetton AD, Crosbie EJ. Metabolomic biomarkers for detection, prognosis and identifying recurrence in endometrial Cancer. *Meta.* 2020;10(8):314.
23. Abul-Husn NS, Kenny EE. Personalized medicine and the power of electronic health records. *Cell.* 2019;177(1):58–69.
24. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. *Bioinform Biol Insights.* 2020;14:1177932219899051.
25. Gajula M. Its time to integrate multi omics data to understand real biology. *Int J Syst Algorithms Appl.* 2012;2:31–4.
26. Tebani A, Afonso C, Marret S, Bekri S. Omics-based strategies in precision medicine: toward a paradigm shift in inborn errors of metabolism investigations. *Int J Mol Sci.* 2016;17(9):1555.
27. Iacobucci I, Wen J, Meggendorfer M, Choi JK, Shi L, Pounds SB, et al. Genomic subtyping and therapeutic targeting of acute erythroleukemia. *Nat Genet.* 2019;51(4):694–704.
28. Soler-Botija C, Gálvez-Montón C, Bayes GA. Epigenetic biomarkers in cardiovascular diseases. *Front Genet.* 2019;10:950.
29. Taha IN, Naba A. Exploring the extracellular matrix in health and disease using proteomics. *Essays Biochem.* 2019;63(3):417–32.
30. Shao Y, Le W. Recent advances and perspectives of metabolomics-based investigations in Parkinsons disease. *Mol Neurodegener.* 2019;14(1):3.
31. Chervitz SA, Deutsch EW, Field D, Parkinson H, Quackenbush J, Rocca-Serra P, et al. Data standards for Omics data: the basis of data sharing and reuse. In: *Bioinformatics for Omics data*. Springer; 2011. p. 31–69.
32. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)toward standards for microarray data. *Nat Genet.* 2001;29(4):365–71.
33. Taylor CF, Paton NW, Lilley KS, Binz PA, Julian RK, Jones AR, et al. The minimum information about a proteomics experiment (MIAPE). *Nat Biotechnol.* 2007;25(8):887–93.
34. Kahl G. Minimum information about a high-throughput nucleotide sequencing experiment (MINSEQE). The dictionary of genomics, transcriptomics and proteomics. Weinheim: Wiley-VCH Verlag GmbH & Co KGaA; 2015.
35. Wurcel V, Cicchetti A, Garrison L, Kip MM, Koffijberg H, Kolbe A, et al. The value of diagnostic information in personalised healthcare: a comprehensive concept to facilitate bringing this technology into healthcare systems. *Public Health Genomics.* 2019;22(1-2):8–15.
36. Ahmed Z. Practicing precision medicine with intelligently integrative clinical and multi-omics data analysis. *Hum Genomics.* 2020;14(1):1–5.
37. Zampieri G, Vijayakumar S, Yaneske E, Angione C. Machine and deep learning meet genome-scale metabolic modeling. *PLoS Comput Biol.* 2019;15(7):e1007084.
38. Voillet V, Besse P, Liaubet L, San Cristobal M, González I. Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework. *BMC Bioinform.* 2016;17(1):1–16.
39. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials—a practical guide with flowcharts. *BMC Med Res Methodol.* 2017;17(1):162.
40. Liu L, Nevo D, Nishihara R, Cao Y, Song M, Twombly TS, et al. Utility of inverse probability weighting in molecular pathological epidemiology. *Eur J Epidemiol.* 2018;33(4):381–92.
41. Malan L, Smuts CM, Baumgartner J, Ricci C. Missing data imputation via the expectation-maximization algorithm can improve principal component analysis aimed at deriving biomarker profiles and dietary patterns. *Nutr Res.* 2020;75:67–76.
42. Zhang Z. Multiple imputation with multivariate imputation by chained equation (MICE) package. *Ann Transl Med.* 2016;4(2):30.
43. Khalid S, Khalil T, Nasreen S. A survey of feature selection and feature extraction techniques in machine learning. In: *2014 science and information conference*. IEEE; 2014. p. 372–8.
44. Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, Molter C, et al. A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Trans Comput Biol Bioinform.* 2012;9(4):1106–19.
45. Vergara JR, Estévez PA. A review of feature selection methods based on mutual information. *Neural Comput Appl.* 2014;24(1):175–86.
46. Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. *Adv Bioinformat.* 2015;198363:1–13.
47. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell.* 2005;27(8):1226–38.
48. Almugren N, Alshamlan H. A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. *IEEE Access.* 2019;7:78533–48.
49. Pal M, Foody GM. Feature selection for classification of hyperspectral data by SVM. *IEEE Trans Geosci Remote Sens.* 2010;48(5):2297–307.
50. Yang L, Xu Z. Feature extraction by PCA and diagnosis of breast tumors using SVM with DE-based

- parameter tuning. *Int J Mach Learn Cybern.* 2019;10(3):591–601.
51. Thankaswamy-Kosalai S, Sen P, Nookaew I. Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. *Genomics.* 2017;109(3-4):186–91.
 52. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics.* 2005;21(9):1859–75.
 53. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics.* 2009;25(14):1754–60.
 54. Sugimoto M, Kawakami M, Robert M, Soga T, Tomita M. Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis. *Curr Bioinforma.* 2012;7(1):96–108.
 55. Cleveland WS, Devlin SJ. Locally weighted regression: an approach to regression analysis by local fitting. *J Am Stat Assoc.* 1988;83(403):596–610.
 56. Corey KM, Kashyap S, Lorenzi E, Lagoo-Deenadayalan SA, Heller K, Whalen K, et al. Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): a retrospective, single-site study. *PLoS Med.* 2018;15(11):e1002701.
 57. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci Rep.* 2016;6(1):1–10.
 58. Kate RJ, Pearce N, Mazumdar D, Nilakantan V. Continual prediction from EHR data for inpatient acute kidney injury. *arXiv preprint arXiv:190210228.* 2019.
 59. Gupta M, Phan TLT, Bunnell T, Beheshti R. Obesity prediction with EHR data: a deep learning approach with interpretable elements. *arXiv.* 2019;p. arXiv–1912.
 60. Lee JM, Hauskrecht M. Multi-scale temporal memory for clinical event time-series prediction. In: International conference on artificial intelligence in medicine. Springer; 2020. p. 313–24.
 61. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. *Sci Rep.* 2018;8(1):1–12.
 62. Chen Y, Li Y, Narayan R, Subramanian A, Xie X. Gene expression inference with deep learning. *Bioinformatics.* 2016;32(12):1832–9.
 63. Chen R, Yang L, Goodison S, Sun Y. Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data. *Bioinformatics.* 2020;36(5):1476–83.
 64. Bruno P, Calimeri F, Kitanidis AS, De Momi E. Data reduction and data visualization for automatic diagnosis using gene expression and clinical data. *Artif Intell Med.* 2020;107:101884.
 65. Thomas J, Thomas S, Sael L. DP-miRNA: an improved prediction of precursor microRNA using deep learning model. In: 2017 IEEE international conference on big data and smart computing (BigComp). IEEE; 2017. p. 96–9.
 66. Bobak CA, Titus AJ, Hill JE. Comparison of common machine learning models for classification of tuberculosis using transcriptional biomarkers from integrated datasets. *Appl Soft Comput.* 2019;74:264–73.
 67. Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.* 2016;44(11):e107.
 68. Yin Q, Wu M, Liu Q, Lv H, Jiang R. DeepHistone: a deep learning approach to predicting histone modifications. *BMC Genomics.* 2019;20(2):11–23.
 69. Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput Biol.* 2017;13(1):e1005324.
 70. Liang CA, Chen L, Wahed A, Nguyen AN. Proteomics analysis of FLT3-ITD mutation in acute myeloid leukemia using deep learning neural network. *Ann Clin Lab Sci.* 2019;49(1):119–26.
 71. Stamate D, Kim M, Proitsi P, Westwood S, Baird A, Nevado-Holgado A, et al. A metabolite-based machine learning approach to diagnose Alzheimer-type dementia in blood: results from the European Medical Information Framework for Alzheimer disease biomarker discovery cohort. *Alzheimer's Dement: Transl Res Clin Interv.* 2019;5(1):933–8.
 72. Muzio G, O'Bray L, Borgwardt K. Biological network analysis with deep learning. *Brief Bioinform.* 2020;22:1515.
 73. Shaban-Nejad A, Michalowski M, Buckeridge DL. Explainability and interpretability: keys to deep medicine. In: Explainable AI in healthcare and medicine. Springer; 2021. p. 1–10.
 74. Anguita-Ruiz A, Segura-Delgado A, Alcalá R, Aguilera CM, Alcalá-Fdez J. eXplainable Artificial Intelligence (XAI) for the identification of biologically relevant gene expression patterns in longitudinal human studies, insights from obesity research. *PLoS Comput Biol.* 2020;16(4):e1007792.
 75. Park S, Kim YJ, Kim JW, Park JJ, Ryu B, Ha JW. Interpretable prediction of vascular diseases from electronic health records via deep attention networks. In: 18th IEEE international conference on bioinformatics and bioengineering, BIBE 2018. Institute of Electrical and Electronics Engineers; 2018. p. 110–7.
 76. Bruno P, Calimeri F, Kitanidis AS, De Momi E. Understanding automatic diagnosis and classification processes with data visualization. In: 2020 IEEE international conference on human-machine systems (ICHMS), vol. 2020. IEEE. p. 1–6.
 77. Hu H, Xiao A, Zhang S, Li Y, Shi X, Jiang T, et al. DeepHINT: understanding HIV-1 integration via deep learning with attention. *Bioinformatics.* 2019;35(10):1660–7.
 78. Choi E, Bahadori MT, Sun J, Kulas J, Schuetz A, Stewart W. Retain: an interpretable predictive model for healthcare using reverse time attention mechanism.

- In: Advances in neural information processing systems. Curran Associates; 2016. p. 3504–12.
79. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. arXiv preprint arXiv:170402685. 2017.
80. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in neural information processing systems. Curran Associates; 2017. p. 4765–74.
81. Björnsson B, Borrebaeck C, Elander N, Gasslander T, Gawel DR, Gustafsson M, et al. Digital twins to personalize medicine. *Genome Med.* 2020;12(1):1–4.
82. Croatti A, Gabellini M, Montagna S, Ricci A. On the integration of agents and digital twins in healthcare. *J Med Syst.* 2020;44(9):1–8.
83. Karczewski K, Snyder M. Integrative omics for health and disease. *Nat Rev Genet.* 2018;19:299–310.
84. Cannataro M, Guzzi PH, Mazza T, Tradigo G, Veltri P. Preprocessing of mass spectrometry proteomics data on the grid. 18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05); 2005. pp. 549–554.
85. Dhillon A, Ashima S. Machine learning in healthcare data analysis: a survey. *J Biol and Today's World* 8 (2019):1–10.
86. Bugnon LA, Yones C, Milone DH, Stegmayer G. Deep neural architectures for highly imbalanced data in bio-informatics. *IEEE Transactions on Neural Networks and Learning Systems* 31(8):2857–2867
87. Talukder A, Barham C, Li X, Hu H. Interpretation of deep learning in genomics and epigenomics. *Briefings in Bioinformatics* 2021;22(3)
88. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision; 2017. pp. 618–626.
89. Zihni E, Madai VI, Livne M, Galinovic I, Khalil AA, Fiebach JB, et al. Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome. *PLoS ONE* 2020;15(4): e0231166.
90. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. Summit on translational bioinformatics, 2010;1–5.



Artificial Intelligence in Evidence-Based Medicine

17

On the Verge of Breakthrough

Artur J. Nowak

Contents

Introduction	256
From PICO Questions to Systematic Reviews	257
Automation of Systematic Reviews	258
Development of Search Strategies	258
Screening	259
Data Extraction	261
New Types of Evidence	262
Making Data More FAIR	262
Other Sources of Data	263
Improving Shared Decision-Making	264
Summary	264
Cross-References	265
References	265

Abstract

There are three pillars of evidence-based medicine (EBM): the evidence itself (e.g., data from clinical studies), clinical expertise, and patient values. EBM is therefore a systematic approach to decision-making that integrates

these three inputs. It involves evidence production (design and conducting of clinical studies), synthesis (collecting, appraising, and combining data to answer clinical questions), implementation (e.g., through clinical practice guidelines based on these syntheses), and evaluation (monitoring the quality of care, including adherence to evidence-based recommendations).

EBM faces many challenges that artificial intelligence can help solve. It can help detect research gaps and avoid funding redundant studies. It can expedite the evidence synthesis process, which currently is slow and costly, leading to outdated and incomplete evidence being used in the decision-making processes.

Supplementary Information: The online version of this chapter (https://doi.org/10.1007/978-3-030-64573-1_43) contains supplementary material, which is available to authorized users.

A. J. Nowak (✉)
Evidence Prime, Krakow, Poland
e-mail: artur.nowak@evidenceprime.com

AI can help engage patients and elicit values (e.g., chatbot-based decision aids) as well as provide coordinated care for patients with multimorbidities.

However, improperly implementing AI can also exacerbate problems in EBM. For instance, if AI-enabled decision support systems fail to incorporate patient values, a return to a model of medicine characterized by low patient autonomy is possible – only this time with a computer in charge.

Keywords

Evidence-based medicine · Natural language processing · Systematic reviews · Screening · Data extraction · Living systematic reviews · Shared decision-making · Clinical practice guidelines · Rapid-learning health systems

Introduction

The term “evidence-based medicine” (EBM) was coined in 1990 at the McMaster University to describe the new approach to teaching the practice of medicine. However, the roots of EBM date back much further, at least to the mid-nineteenth century. Through the 1970s and 1980s, researchers such as Archie Cochrane and David M. Eddy debunked claims of effectiveness of many interventions by pointing out that they didn’t stand a test of a randomized clinical trial. This data-driven approach, in which arguments were based on results of carefully designed experiments (evidence) rather than authority, was a revelation – in large part by showing how often the “authority” was wrong.

The emphasis on cold hard facts was foundational for EBM, and it remains its most visible facet. However, the evidence itself is only one of the three pillars of EBM, the remaining two being patient values and preferences, and clinical expertise. Therefore, evidence-based medicine is defined as “a systematic approach to clinical problem solving, which allows the integration of the best available research evidence with clinical expertise and patient values” [1]. The patient

values and preferences should be considered both on the population level, during the development of clinical practice guidelines, and on the individual level by engaging patients in shared decision-making.

Today, EBM is the widely accepted paradigm of medical practice. Other disciplines, such as management or software engineering, are also adopting evidence-based approaches. However, the implementation of EBM still faces many challenges. According to [2], the effectiveness of healthcare has reached a plateau, in which only 60% of care is in line with guidelines (evidence-based or not), 30% is some form of waste of low value, and 10% is harm. This subpar performance has persisted at least for the last three decades, which is since EBM’s formal inception.

Over the years, many criticisms of EBM have been published [3]. As EBM promotes the use of randomized controlled trials, which often have stringent enrollment criteria, it has been pointed out that the effect estimates may be biased [4]. Moreover, integrating these population-level results with personal clinical expertise and applying them to individual patients is difficult. Some doctors have seen EBM as a threat to their autonomy and their relations with patients. Another limitation that has been cited is that EBM requires thousands of (expensive) clinical trials that answer narrow questions. Finally, a common criticism was that the evidence used for decision-making is outdated, as there is a significant lag between study publication and incorporating it in recommendations.

Artificial Intelligence (AI) becomes an increasingly important tool for tackling these issues. Employing AI to expedite the evidence synthesis process not only opens new opportunities but is quickly becoming a necessity due to the exponential growth of scientific output. Researchers also hope to leverage AI’s analytical capabilities to integrate evidence from previously untapped sources, such as medical sensors, electronic health records (EHRs), and pharmacovigilance databases.

The original vision of EBM included trained clinicians who had tools and skills to incorporate the best available evidence with their internal expertise and patient’s values and preferences.

The implementation remains challenging, as this task requires integrating objective, population-level data with the personal needs of a patient. AI is a transformative force capable of solving these issues. Improperly used, it can also create problems of its own [5].

From PICO Questions to Systematic Reviews

The building blocks of EBM are clinical questions. The most common type is the PICO (Population Intervention, Comparison, Outcomes) question, which compares the effects of two interventions applied to a health problem or population as measured by a set of outcomes (e.g., mortality, average blood pressure). Other mnemonics are used for different health problems, such as PECO for environmental exposures. The first step in answering these questions is identifying all relevant evidence, preferably through the systematic review process.

Systematic Reviews (SR) offer a comprehensive and impartial synthesis of all human knowledge on a given topic, resulting from a standardized, transparent, and meticulous process. The primary goal is to maximize sensitivity (recall) – find all relevant studies. At the highest level of generality, the process can be broken down into eight steps:

1. A review protocol is created, which documents the basic assumptions which will guide researchers. Criteria for the inclusion and exclusion of evidence are defined. Both the search sources to be used (publication databases, clinical trial registries, other reviews) and the search strategy for these sources are described. The protocol also contains a description of the methodology for collecting data for review and analysis.
2. Searches are conducted in indicated databases and other sources using the strategy defined in the protocol. The search strategies are expressed as Boolean queries using database-specific syntaxes. The net is cast wide to maximize the sensitivity, so the search often returns thousands

of results. In general, databases contain only bibliographic metadata, titles, and abstracts. Full texts of papers need to be fetched separately.

3. The results are deduplicated and screened by applying inclusion and exclusion criteria to titles and abstracts. Ideally, two reviewers perform this step independently to reduce the number of false negatives (relevant documents excluded by mistake).
4. Full texts of references that pass the first screening stage are then retrieved. Reviewers screen the records by reading the whole text and applying more detailed inclusion criteria – especially concerning information not present in the abstracts.
5. Data is extracted from text, tables, and plots into a structured format (e.g., a spreadsheet), which can be used for quantitative analysis.
6. The results from individual studies are combined, preferably using meta-analysis. Increasingly, questions assess more than two interventions at the same time by leveraging network meta-analysis.
7. A critical appraisal of the evidence is carried out. The most recognized approach used at this stage is the methodology developed by the GRADE Working Group (Grading of Recommendations, Assessment, Development and Evaluations).
8. Due to the continuous progress of medical knowledge (e.g., in cardiology, there are 150 new studies published every week), updating is a critical but often neglected element of the process.

In 2014 alone, MEDLINE indexed over 8,000 systematic reviews. According to some estimates, almost 14,000 are published annually (this does not include HTA reports and internal reports created by the industry). The primary source of costs during the creation of systematic reviews is human labor. Therefore, the total cost depends on the time it takes to create a review, which varies from 6 to 24 months, subject to the scope and complexity. According to the 2013 estimate based on UK rates, a review's total cost is approximately \$100,000 (often just for one question).

These costs grow dynamically every year, which is not surprising given the increase in the number of articles published – over 800,000 papers are indexed annually by MEDLINE; nearly 30,000 clinical trials are registered with clinicaltrials.gov.

Moreover, due to the time and cost required to create or update a systematic review, they cannot keep up with the newly published studies. According to [6], 7% of reviews are outdated at the time of their publication, and after two years, about 23% of outdated reviews will contain conclusions inconsistent with the new medical knowledge.

The situation of the COVID-19 pandemic further amplified both these problems. On the one hand, it required urgent guidance, which would adapt to the changing situation. On the other, an explosion of published research of mixed quality made it difficult to find all relevant studies on time.

Living Systematic Reviews were postulated [7] to address the lack of currency of systematic reviews. These reviews are continually updated, incorporating data from new studies as soon as they appear. However, the help of AI is needed to make this process sustainable.

Automation of Systematic Reviews

As early as 2006 [8], machine learning was applied to selecting studies (screening step in the systematic review process). In 2014, a survey [9] was published that presented examples of 10 automation technologies and discussed future research. International Collaboration for the Automation of Systematic Reviews (ICASR) was established in 2015 to foster collaboration between groups working on review automation. The group holds annual meetings to discuss progress and identify the next challenges of applying AI to the automation of systematic reviews [10].

The primary goal of the automation efforts is increasing the speed and reducing the workload. A recent report describes a systematic review completed in just two weeks using automation tools [11]. However, the introduction of advanced automation tools can also modernize the process

and improve quality. Many of the tasks in the process are tedious and error prone, especially without adequate tool support. Therefore, AI can serve as an additional pair of eyes, which can catch human errors. This application becomes even more relevant in the context of rapid systematic reviews [12]. Such studies are required in situations requiring an urgent response, such as chemical spills or disease outbreaks. One example of the latter is the COVID-19 pandemic [13].

Today, 22 software tools use machine learning in some capacity [14]. They range from small utilities that help with a single aspect of the systematic review process to integrated platforms that facilitate the process from start to finish.

Development of Search Strategies

A Web search engine user is interested primarily in finding the most relevant result on the first page of the results. For this reason, the frequently optimized metric for this problem is P@5 (“precision at 5”) – the proportion of relevant hits within the top 5 results returned by the engine. Conversely, systematic reviews aim to find all relevant studies, achieving perfect recall (sensitivity). For this reason, information retrieval techniques popularized by Web search engines are not directly applicable to systematic reviews. Therefore, the searches are currently performed using manually constructed Boolean queries – they are logic sentences combined with AND, OR, and NOT operators (detecting a term’s presence or absence).

The existing tools that aid in searching can be broadly classified as:

1. Increasing sensitivity of the search by adding words to a query. This is done based on the linguistic analysis of intermediate results (e.g., SWIFT-Review, BibExcel). There are also systems supporting the selection of keywords and keywords from dictionaries such as MeSH (PubMed PubReminer, Yale MeSH Analyzer).
2. Increasing precision by replacing individual words with word combinations (e.g., TerMine) or clustering results and identifying features to exclude some hits (AntConc) safely.

3. Supporting search in many databases by translating between syntaxes (e.g., Polyglot).
4. Automatically generating a search strategy based on a set of seed articles – records identified as relevant by experts.
5. Automated tagging of study designs (RobotSearch) or PICO elements (e.g., PICO Portal, Trip Database, DOC Search) to index references.

The last type of tools come closest to fully leverage the potential of AI for addressing the problem of search. Still, they are currently unable to produce search strategies in the traditional (Boolean) form. The current search strategy format has undeniable advantages: repeatability, syntax compatibility with existing databases, and the availability of tools translating between syntaxes. Moreover, it is possible to perform them on practically all databases in the world (which is essential when searching the so-called gray literature).

Whether the advantages of AI-assisted search will outweigh the problems related to interpretability and portability remains to be seen. The latter problem could be mitigated by better data integration. The reproducibility of searches based on machine learning models' results could be addressed by careful versioning. Some projects are exploring hybrid approaches, in which Boolean search strategies are generated retrospectively to match the results coming from an AI-enhanced search.

Finally, search strategies also need to be monitored. Changes in the corpus contents (the so-called topic drift, e.g., the appearance of a new drug) may result in the current search strategy not finding all relevant documents anymore. It is another use for automation that is currently being explored. Some examples include walking the citation graph (i.e., automatically screening documents that cite studies included in the review) to check if the current search strategy captures them.

Screening

The very need for screening comes from the previous step's specifics. As discussed previously, the execution of a strategy is limited to

determining whether certain words or concepts (e.g., MeSH) appear in the document or not. For this reason, it is not possible to narrow down search results by adding the condition “and there is no word” in a safe manner (i.e., without loss of sensitivity), for fear of a situation where the word occurs in a context different than intended. For example, if one looks for randomized studies, they cannot add the term “NOT observational” to the query, as this term may appear at least in part describing previous studies.

Consequently, search results are characterized by high sensitivity (a small number of false negatives) but low precision (a large number of false positives). The task of separating the wheat from the chaff, that is, eliminating articles that do not meet the inclusion criteria, lies with those who carry out screening of abstracts and then full texts.

Machine learning helps in this process by learning from screening decisions and applying them to yet unscreened references. Typically, the documents are transformed into a review-specific representation, most commonly TF-IDF vectors. Interestingly, document representations resulting from deep learning language models (e.g., BERT) fail to outperform this simpler approach consistently. The references for which decisions are available are used as a training set. The model then predicts the score for the yet unscreened references. This score can be used to rank the records, so the reviewers look at the most promising records first.

This process is most effective when it is interactive, that is, the model retrains continuously on new decisions from users and updates the prioritized list. This scenario is a special case of active learning. The algorithm can query a user (called “teacher” or “oracle”) to label examples that could define the decision boundary. This method proved to be very effective for another text classification task that requires high sensitivity: sifting through legal evidence in eDiscovery [15]. Screening tools that employ some form of active learning include Abstrackr, RobotAnalyst, SWIFT-Active Screener, litstream, DistillerSR, EPPI-Reviewer, Laser AI, ASReview [16].

In the active learning scenario, it is also possible to estimate the current level of recall [17], so the

users can decide when to stop screening, automatically excluding all remaining documents. Such functionality is available, for example, in SWIFT-Active Screener, litstream, and DistillerSR. However, the systematic review community hasn't established a widely accepted stopping criterion yet [18]. Many groups are currently evaluating this functionality to test its robustness across a number of reviews.

Systematic review methodologies (e.g., Cochrane handbook) recommend that screening is performed independently by two researchers to reduce the possibility of human error. If the two reviewers don't agree, they either resolve the conflict in a discussion, or a third reviewer acts as an adjudicator. A common approach for the titles and abstracts screening phase is to include a reference to the next stage if at least one reviewer decides to mark it as relevant.

As many projects lack resources to perform double screening on all the references, a machine learning model can fill this gap and act as a second (or third) screener. To avoid bias, this method is more appropriate for situations where reviewers perform screening without machine assistance. Their decisions can then be split into two folds. The first one is used to train a model, which is then run on the second fold. The folds are then swapped, and the process is repeated. As a result, machine-provided scores are available for all the records and are compared against human decisions to spot any errors.

Classifiers trained on previous screening decisions are also used during review updates and literature surveillance as part of living systematic reviews. It remains to be seen how robust these classifiers are over long periods, for example, to phenomena such as the topic drift mentioned above. During training, the generalizability of these classifiers is commonly evaluated on a held-out set that was chosen randomly, without taking the publication date into account. It may be then a case that if new documents come from a different distribution (e.g., new words for novel interventions appear), the accuracy of these models will decrease. This underlines the importance of monitoring search strategies for their effectiveness.

While most titles and abstracts are available in English, it is not true for full texts. It was especially evident during the COVID-19 pandemic when many early studies were published in Chinese. AI can help here as well, thanks to the high maturity of machine translation. While the final interpretation of results during the data extraction step will probably require a human translator, automatic translation can streamline the screening process, decreasing the number of papers that need to be translated by a human.

While automation of screening has already achieved widespread adoption in several tools and has been successfully applied in many projects, some challenges remain. The major ones connect to the need to achieve almost perfect sensitivity with precision at such a level to reduce the time needed for screening significantly. This task is challenging mainly due to the wide variety of topics of systematic reviews. A serious problem is a large variance in the accuracy of the current tools, depending on the nature of the review [19]. To illustrate this problem, one can think about the levels of autonomy of vehicles. AI models are already capable of driving cars in well-known and controlled environments (e.g., highways), but extending this autonomy level to all roads is a much harder task.

This problem is exacerbated by the lack of interpretability of the results of the model. So far, there is no way to visualize the results of automatic screening, which would allow users to make sure that it is working correctly. A particular challenge here is the number of results for which clear representation (tens of thousands) must be constructed. Moreover, the larger the current NLP models become, the harder they are to understand by researchers and end-users. Without insight into how the model represents the documents and makes decisions, it is difficult for the users to determine if the model is likely to make errors on their data. To extend the self-driving cars analogy, users don't know what driving conditions can make the system fail.

Since most current systems learn only from previous decisions, they struggle with reviews that have very few positive examples (included studies). In an extreme case, systematic reviews

that aim to identify research gaps can find no studies to include. In such situations, classifiers based on previous screening decisions simply cannot be trained. A promising direction for solving this problem is the use of the new generation of few-shot models that could use the search strategy and information on inclusion and exclusion criteria in addition to screened records.

Incorporating the inclusion and exclusion criteria as an input of the model can also address indirect evidence. Sometimes, studies that directly answer the question at hand are scarce or unavailable. Therefore, researchers need to decide what level of indirectness in the population (e.g., different or mixed age groups), interventions (e.g., larger dose), or outcomes (e.g., follow-up length) is acceptable. These nuances may be difficult to capture by a classifier if labels are binary – a study is either included or excluded. Therefore, models that can represent the study text and the query (inclusion criteria) are a promising direction for the next step in screening automation.

Data Extraction

At the highest level of generality, data extraction consists of converting data in semi- or unstructured form (article text, figures, tables) into a fully structured form (data extraction form). Each systematic review can define its unique fields, which will be extracted. Cochrane recommends that 36 categories be considered when creating the data extraction form and that an additional 7 items should be included depending on the review type. This number may be higher depending on the specifics of the question and the types of studies included. For example, teams using the ROBINS-I tool to assess the risk of bias in observational studies will need far more fields than Cochrane recommend. The fields also create hierarchical relationships and appear repeatedly: for example, study arms will be extracted separately for each intervention, measurement results and units will appear separately for each outcome, etc.

This poses difficulties: a list of the fields that will appear most frequently should be drawn up

and defined in such a way (i.e., what information is to be found in them) that it meets the needs of the whole community. Still, every review may feature some fields for which the model was not trained. Therefore, while currently most of the automation tools frame this problem as named entity recognition or classification (if the list of values is controlled), it would be optimal to leverage QA models to allow open-ended (zero-shot) data extraction.

The technical challenge is not only the number of fields but also the rich structure of the annotation. It is not enough to state that a number appearing in the text indicates the size of the study group – it is also necessary to assign this value to a specific group, the description of which appears perhaps in the latter part of the text. Besides, extracted strings of text may be discontinuous and overlapping: for example, the outcomes “bleeding (total)” and “bleeding (serious)” may be represented in the text by the same word “bleeding,” while in the latter case, the determination of the severity appears only a few words later.

Moreover, while some of the data points can be easily found in the study text, others are present only in tables or need to be read from plots. Sometimes the extraction involves some level of interpretation and assessment. Was the randomization performed correctly? When the study authors report mortality in the study, does it include or exclude patients lost to follow-up? These examples may suggest that the goal of AI deployment for this problem is to semiautomate the work – augment abilities of human reviewers, so they can concentrate on critical appraisal of the studies.

To sum up, for a typical question with nine endpoints, for which 19 studies were identified, it is necessary to extract an average of 80 data points per study, that is, over 1,500 per question. If, as recommended, the work is carried out independently by two people, this translates to about 80 working hours, not including discussions about discrepancies. It is a strong argument for the need to improve this process, which at the moment is often performed using spreadsheets.

Time-consuming data extraction also hinders the inclusion of other data sources in reviews. The

European Medicines Agency committed to publishing clinical data, including Clinical Study Reports (CRS), of all products submitted for the market authorization procedure. Although the publication was temporarily suspended in 2019 (with the exception of COVID-19 medicines), this data source is expected to grow in the future. These documents, albeit challenging to process, will be an invaluable source of knowledge for groups analyzing research results.

A statistical model, which can recognize information in the text with great accuracy, is only half of the success. Similar to how it was discussed for screening, the results need to be interpretable. Otherwise, one difficult-to-navigate format (multi-page article text) will be changed to another – a list of hundreds of unordered, extracted text fragments, the correctness of which cannot be judged.

A systematic review of automation tools from 2015 [20] identified a number of systems, which were capable of identifying sentences or individual concepts with high accuracy for simple data types, such as sex or age (F1, a harmonic mean of precision and recall, is the most common metric used to compare systems). In total, of the 52 fields recommended by at least one of the leading standards, only 23 are currently extracted by at least one tool (44%). RobotReviewer, which, in its current version, claims that it can mark all 23 fields, leading the way in this respect.

In 2018, the Systematic Review Information Extraction Track was organized as part of the Text Analysis Conference to develop and evaluate automated data extraction approaches that can assist in the systematic reviews of environmental agents [21]. The track results highlighted the advantages of using deep learning models paired with contextual embeddings (ELMo). This direction has recently seen significant development in the field of Natural Language Processing with models based on the Transformer architecture [22]. It is then expected that the tools for automation of data extraction will see significant improvements. A living systematic review has been started to monitor the progress of the field [23].

New Types of Evidence

The description above concentrates on semi-automation of evidence synthesis from clinical study reports. Two questions may arise. First, why are the results reported as text, which is difficult to find and process? Second, what about other sources of data?

Making Data More FAIR

The first question relates to the idea of FAIR (Findable, Accessible, Interoperable, Reusable) data. Data that is *findable* is indexed in a searchable database using terms that allow its retrieval. Study authors would need to use medical ontologies to map their work to PICO components that systematic review authors use in their inclusion criteria. Precise and consistent coding of these complex concepts by all researchers would be required. The data would then need to be easily downloadable using well-known methods to be also *accessible*. Additionally, the data would need to use standardized formats to be *interoperable*. Finally, it would need to clearly define the semantics of fields and usage license to be *reusable*.

Multiple efforts were made, also by study funders, to improve FAIRness of clinical data. The most extensive database is [clinicaltrials.gov](#), established in 2000 and currently storing information on more than 360,000 studies. However, the vast majority of these records contain only the most basic attributes. Many of the completed trials were not updated with their results. In effect, the registries such as [clinicaltrials.gov](#) are currently used as a source of documents in the screening process, but the study reports need to be retrieved from literature databases as well. To improve the situation, a 2016 federal regulation was introduced that requires all NIH-funded trials to submit summary results information to [clinicaltrials.gov](#). There are also initiatives from evidence producers [24] and synthesizers [25] to develop more advanced formats that adhere to the FAIR principles.

The role of AI in this process is twofold. Firstly, it is a major driver of this change, as improving FAIRness of data makes it easier to feed into AI systems. Secondly, the NLP models developed for systematic review automation can be deployed to help researchers convert their results into standardized formats. One example of such applications would be a system that analyzes a submitted manuscript to prepopulate FAIR representation automatically. Additionally, such tools could be applied retroactively – with human supervision – to produce machine-readable versions of existing papers.

From the resource use perspective, the situation in which authors describe their findings in text so the other researchers can then extract them back into a tabular format seems like a waste. However, the study reports contain much more information than just the effect estimates, which sometimes needs to be read between the lines. The manuscripts are here to stay, as they are easier to write, peer-review, and read than trying to encode the equivalent information in a structured, machine-readable format. Additionally, moving to a more efficient medium for communicating research results requires a critical mass and coordination of researchers worldwide. It is challenging to enforce the standards for knowledge representation in such a distributed scenario. However, the current efforts clearly show that both researchers, funders, and publishers recognize the importance of improving the reporting standards and making data sharing easier. The COVID-19 pandemic once again has a chance to be the catalyst for improved coordination among these actors.

Other Sources of Data

The defining quality of AI is its ability to process large amounts of multidimensional data. It is capable of discovering patterns and detecting correlations between numbers of variables that far exceed human capacity. EBM can leverage these traits to tap into data that was previously impossible to process and analyze [26].

These data sources include electronic health records (EHR), medical claims databases, pharmacovigilance registries, medical sensors, social networks, patient-reported data, and omics (genomics, proteomics, etc.). Sometimes they are called “real-world data” or “real-world evidence” to underline the fact that they come from outside of controlled conditions of clinical trials. However, some find this term misleading by suggesting that other evidence is not “real” [27]. The “real-world” retrospective studies are also tested to see if they can replicate the results of randomized controlled trials on efficacy [28]. However, the most promising direction seems to be integrating different types of evidence, extending the clinical trials’ findings with additional insight from populations that are not typically included.

Firstly, AI allows transforming these streams of data into representations suitable for further analysis. For instance, NLP models are used to extract data from text entries in EHRs and automatically assign medical codes. Similarly, sentiment analysis models are applied to social media posts and patient discussion forums. Machine learning models also help tackle data quality issues by automatically detecting outliers or linking different records for the same patient.

Secondly, machine learning models are applied to connect the dots and find correlations in these vast amounts of data. For instance, studies now explore how patient-reported data can be used [29]. It is natural to expect that advanced data science methods will become indispensable as the number of data points collected increases. However, one needs to remember the old adage that “correlation is not causation.” All hypotheses generated by AI must be evaluated with the same rigor that is fundamental for evidence-based medicine.

Expanding the scope of collected data can also help close the feedback loop in the EBM ecosystem. EHR and patient-reported data can help track quality measures and progress in implementing clinical practice guidelines. Examples of measured outcomes include the rate of postoperative complications, hospital-acquired infections, or

reports on the quality of life. However, these data points cannot be interpreted as is, as the results are subject to many factors, such as patient health status. Machine learning models can help analyze many data points related to the patients to produce more accurate risk-adjusted estimates that correct for these confounders.

Moreover, integrating new evidence sources can help identify research gaps and better prioritize research funds. In general, applying evidence-based practices to the research itself has been postulated [30], for example, by the EVBRES (EVidence-Based RESearch) consortium. Meta-research (i.e., research on research) studies show that the evidence production (including planning and design of new studies) isn't done systematically and transparently. As a result, too many redundant studies are conducted – leading to a waste of time and resources. What is more critical, duplicated studies mean that patients unnecessarily receive placebo or suboptimal treatment, which leads to a waste of health and life. AI for automation of evidence synthesis can help more efficiently detect redundant studies at early stages through its methods for expediting the identification of existing evidence.

Improving Shared Decision-Making

One of the pillars of EBM is to include patient values and preferences in the process of decision-making. GRADE Evidence to Decision framework [31] explicitly considers them one of its criteria for recommendation development. This process is likely to benefit from additional information sources on patient values and behavior (e.g., adherence) as discussed above. However, these recommendations are designed to be further contextualized to the needs of an individual through a process of shared decision-making (SDM). SDM is a process in which patient and physician together take part in the decision-making and agree on treatment decisions. Health care providers facilitate the discussion by presenting possible options (along with their benefits and harms) to patients. The final decision reflects patient preferences as well as their lifestyle and religious and cultural specifics.

The United States Preventive Services Task Force (USPSTF) recommend that patients and physicians participate in shared decision-making. However, patient involvement remains low [32]. Most often, SDM is left to the initiative of the physician. Despite the efforts to standardize this process, the integration of patient values in Clinical Decision Support Systems (CDSSs) is rare [33]. In other words, the technology that is the primary channel of supplying evidence during the clinical encounter is not providing support for patient input.

This issue could be addressed by a new generation of AI-enabled decision aids that provide better user experience, such as chatbots. Additionally, CDSSs can become value-flexible [5] by using decision models and processes [33] that incorporate stakeholder inputs. They can also benefit from including more data in the process – dietary restrictions and genomics, but also experiences of other patients. The latter can be retrieved by applying AI to analyze patient forums [34].

An increasing number of patients suffer from multiple diseases. The management of multimorbid patients is challenging, as the number of risk factors increases with the number of clinical conditions. Unfortunately, few clinical practice guidelines address multimorbidity. Hence, EBM recommendations may lead to conflicts and adverse events, such as interactions between drugs. Coordinated care also involves consideration of individual values and lifestyle characteristics. Guidelines are being converted to machine-interpretable formats to allow CDSSs to reason about their overlaps [35]. Researchers are also working on AI models that can automatically solve the conflicts between the recommendations [36]. If these efforts are successful, a new generation of CDSSs can emerge that genuinely deliver on the promise of personalized medicine.

Summary

In 2007, rapid-learning health systems were proposed [37] to continuously improve the quality and cost-effectiveness by analyzing information collected “on the ground,” for example, from databases of electronic patient data. This idea

has been since implemented in a number of health systems worldwide. Today, AI enables more efficient and faster evidence synthesis, bringing the idea of living systematic reviews closer to the reality. At the same time, use of machine learning enables collecting, curating, and analyzing much larger amounts of data. Together, these innovations can be used to build deep-learning health systems [38].

Deep-learning health systems will put the current, best available evidence at the fingertips of doctors and help them make decisions by leveraging AI-enabled decision support systems. The information on the clinical outcomes would then be fed back into the system, allowing it to learn – very much the same way as deep neural networks learn through backpropagation. At the same time, value-insensitive decision support systems present a potential threat to shared decision-making, because the individual patient's values and preferences are considered only as an afterthought. Instead, AI needs to be used in a way that embraces deep medicine [39] – in the way that makes medicine human again.

Cross-References

- [AIM and the Patient's Perspective](#)
- [AIM in Electronic Health Records \(EHRs\)](#)

References

1. Haynes RB, Sackett DL, Richardson WS, Rosenberg W, Langley GR. Evidence-based medicine: how to practice & teach EBM. *Can Med Assoc J*. 1997;157(6):788.
2. Braithwaite J, Glasziou P, Westbrook J. The three numbers you need to know about healthcare: the 60-30-10 challenge. *BMC Med*. 2020;18(1):102.
3. Cohen AM, Stavri PZ, Hersh WR. A categorization and analysis of the criticisms of evidence-based medicine. *Int J Med Inf*. 2004;73(1):35–43.
4. Krauss A. Why all randomised controlled trials produce biased results. *Ann Med*. 2018;50(4):312–22.
5. McDougall RJ. Computer knows best? The need for value-flexibility in medical AI. *J Med Ethics*. 2019;45(3):156–60.
6. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? A Survival Analysis. *Ann Intern Med*. 2007;147(4):224.
7. Elliott JH, Synnot A, Turner T, Simmonds M, Akl EA, McDonald S, et al. Living systematic review: 1. Introduction – the why, what, when, and how. *J Clin Epidemiol*. 2017;91:23–30.
8. Cohen AM, Hersh WR, Peterson K, Yen P-Y. Reducing workload in systematic review preparation using automated citation classification. *J Am Med Inform Assoc JAMIA*. 2005/12/15 ed. 2006;13(2):206–19.
9. Tsafnat G, Glasziou P, Choong MK, Dunn A, Galgani F, Coiera E. Systematic review automation technologies. *Syst Rev*. 2014;3(1):74.
10. O'Connor AM, Glasziou P, Taylor M, Thomas J, Spijker R, Wolfe MS. A focus on cross-purpose tools, automated recognition of study design in multiple disciplines, and evaluation of automation tools: a summary of significant discussions at the fourth meeting of the international collaboration for automation of systematic reviews (ICASR). *Syst Rev*. 2020;9(1):100.
11. Clark J, Glasziou P, Del Mar C, Bannach-Brown A, Stehlik P, Scott AM. A full systematic review was completed in 2 weeks using automation tools: a case study. *J Clin Epidemiol*. 2020;121:81–90.
12. Schünemann HJ, Moja L. Reviews: rapid! Rapid! Rapid! ...and systematic. *Syst Rev*. 2015;4(1):4. 2046-4053-4-4
13. Chu DK, Akl EA, Duda S, Solo K, Yaacoub S, Schünemann HJ, et al. Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: a systematic review and meta-analysis. *Lancet*. 2020;395(10242):1973–87.
14. Systematic Review Toolbox [Internet]. [cited 2020 Dec 30]. Available from: <http://systematicreviewtools.com/>
15. Grossman MR, Cormack GV. Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Rich JL Tech*. 2010;17:1.
16. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Syst Rev*. 2019;8(1):163. s13643-019-1074-9
17. Cormack GV, Grossman MR. Scalability of continuous active learning for reliable high-recall text classification. In: Proceedings of the 25th ACM international conference on information and knowledge management [Internet]. Indianapolis Indiana: ACM; 2016. [cited 2020 Dec 31]. p. 1039–48. <https://doi.org/10.1145/2983323.2983776>.
18. Hamel C, Kelly SE, Thavorn K, Rice DB, Wells GA, Hutton B. An evaluation of DistillerSR's machine learning-based prioritization tool for title/abstract screening – impact on reviewer-relevant outcomes. *BMC Med Res Methodol*. 2020;20(1):256.
19. Gates A, Johnson C, Hartling L. Technology-assisted title and abstract screening for systematic reviews: a retrospective evaluation of the Abstrackr machine learning tool. *Syst Rev*. 2018;7(1):45.
20. Jonnalagadda SR, Goyal P, Huffman MD. Automating data extraction in systematic reviews: a systematic review. *Syst Rev*. 2015;4(1):78.

21. Schmitt C, Walker V, Williams A, Varghese A, Ahmad Y, Rooney A, et al. Overview of the TAC 2018 systematic review information extraction track. TAC; 2018.
22. Cohan A, Feldman S, Beltagy I, Downey D, Weld DS. SPECTER: document-level Representation Learning using Citation-informed Transformers. ArXiv200407180 Cs [Internet]. 2020 [cited 2020 Dec 31]; Available from: <http://arxiv.org/abs/2004.07180>
23. Schmidt L, Olorisade BK, McGuinness LA, Thomas J, Higgins JPT. Data extraction methods for systematic review (semi)automation: a living review protocol. F1000Research. 2020;9:210.
24. Wise J, de Barron AG, Splendiani A, Balali-Mood B, Vasant D, Little E, et al. Implementation and relevance of FAIR data principles in biopharmaceutical R&D. Drug Discov Today. 2019;24(4):933–8.
25. Alper BS, Richardson JE, Lehmann HP, Subbian V. It is time for computable evidence synthesis: the COVID-19 knowledge accelerator initiative. J Am Med Inform Assoc. 2020;27(8):1338–9.
26. Scott I, Cook D, Coiera E. Evidence-based medicine and machine learning: a partnership with a common purpose. BMJ Evid-Based Med. 2020;bmjebm-2020-111379.
27. Schünemann HJ. All evidence is real world evidence – The BMJ [Internet]. [cited 2021 Jan 3]. Available from: <https://blogs.bmjjournals.org/bmjj/2019/03/29/holger-j-schunemann-all-evidence-is-real-world-evidence/>
28. Franklin JM, Patorno E, Desai RJ, Glynn RJ, Martin D, Quinto K, et al. Emulating randomized clinical trials with nonrandomized real-world evidence studies: first results from the RCT DUPLICATE initiative. Circulation. 2020; CIRCULATIONAHA.120.051718.
29. Bédard A, Basagaña X, Anto JM, García-Aymerich J, Devillier P, Arnavielhe S, et al. Mobile technology offers novel insights into the control and treatment of allergic rhinitis: the MASK study. J Allergy Clin Immunol. 2019;144(1):135–143.e6.
30. Robinson KA, Brunnhuber K, Ciliska D, Juhl CB, Christensen R, Lund H. What evidence-based research is and why is it important? J Clin Epidemiol. 2020; S0895435620310957
31. Alonso-Coello P, Schünemann HJ, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. BMJ. 2016;353:i2016.
32. Couët N, Desroches S, Robitaille H, Vaillancourt H, Leblanc A, Turcotte S, et al. Assessments of the extent to which health-care providers involve patients in decision making: a systematic review of studies using the OPTION instrument. Health Expect. 2015;18(4):542–61.
33. Parimbelli E, Wilk S, Kingwell S, Andreev P, Michalowski W. Shared decision-making ontology for a healthcare team executing a workflow, an instantiation for metastatic spinal cord compression management. AMIA Annu Symp Proc AMIA Symp. 2018;2018:877–86.
34. Yang J, Xiao L, Li K. Modelling clinical experience data as an evidence for patient-oriented decision support. BMC Med Inform Decis Mak. 2020;20(S3):138.
35. Bilici E, Despotou G, Arvanitis TN. The use of computer-interpretable clinical guidelines to manage care complexities of patients with multimorbid conditions: a review. Digit Health. 2018;4:205520761880492.
36. Ćyras K, Oliveira T. Resolving conflicts in clinical guidelines using argumentation. ArXiv190207526 Cs [Internet]. 2019 [cited 2021 Jan 5]; Available from: <http://arxiv.org/abs/1902.07526>
37. Etheredge LM. A rapid-learning health system: what would a rapid-learning health system look like, and how might we get there? Health Aff (Millwood). 2007;26(Suppl1):w107–18.
38. Norgeot B, Glicksberg BS, Butte AJ. A call for deep-learning healthcare. Nat Med. 2019;25(1):14–5.
39. Topol EJ. Deep medicine: how artificial intelligence can make healthcare human again. 1st ed. New York: Basic Books; 2019. 1 p.



AIM in Electronic Health Records (EHRs)

18

Yi Guan and Jingchi Jiang

Contents

Introduction	268
The Symbolic-Connectionism Cognitive Architecture	268
The Holistic Framework of Research Work of WI Lab in AIM	271
Data Tier	272
Knowledge Tier	272
Application Tier	273
Selected Research	277
Corpus Construction	277
Information Extraction	277
Knowledge Expansion	277
Clinical Reasoning Model	280
Disease Risk Prediction	282
Conclusion and Future Work	284
References	285

Abstract

Artificial intelligence (AI) in medicine prefers rationalism. Medicine not only imposes higher requirements on the rationality of the AI but also provides a restricted domain for AI that is relatively easy to make breakthroughs. This chapter introduces research work c) in electronic health records (EHRs) carried out by the Web Intelligence Laboratory (WI Lab) under the guidance

of scientific theories such as cognitive science, complex network, and machine learning. To build the entire research work on a solid foundation of cognitive science, the authors first propose the symbolic-connectionism cognitive architecture theory and then explain representative research works such as medical corpus construction, natural language processing (NLP), and information extraction (IE) at the data tier; knowledge representation and medical ontology construction at the knowledge tier; and disease diagnosis, electronic medical record quality control, overtest detection, disease risk prediction, chronic disease management, and dermatosis diagnosis at the application tier. A rationalist

Y. Guan (✉) · J. Jiang
WI (Web Intelligence) Lab, School of Computer Science
and Technology, Harbin, China
e-mail: guanyi@hit.edu.cn

route with the core of automatic knowledge mining from medical big data, medical knowledge representation, and reasoning has been explored in practice.

Keywords

Artificial intelligence in medicine · Electronic health records · Cognitive architecture · Symbolic-connectionism · Knowledge mining · Medical knowledge representation · Medical ontology · Reasoning

Introduction

The development of deep learning not only pushes AI to an unprecedented height but also significantly promotes its applications in the medical field [1]. However, the latter is constrained by its two inherent properties: first, since deep learning techniques originate from a neural network, its processing power is limited to classification, while most of the thinking process in medicine is reasoning [2]. Second, the principles of deep learning have not been revealed, and their effectiveness is not reliable so that their applications in medicine still face many legal and ethical issues.

To delve into AIM research, it must be acknowledged that **AIM prefers rationalism**. It is not just because medicine is a matter of life and death so that the techniques it accepts should always avoid leading patients' condition to uncontrollable or unpredictable stages, but reasoning itself is a topic of AIM. Modeling of cognitive capabilities beyond classification, such as reasoning and decision-making, must be included in the scope of research. On the other hand, medicine provides reasoning with a restricted domain, whose complexity is greatly reduced compared to the open domain.

For AIM research of WI Lab, rationalist route has dual meanings: first, mining medical knowledge to reason is the primary research topic of the laboratory. Second, researchers should base all research on solid cognition science, complex network, and machine learning theories and try to propose white-box methods.

To build work on the solid cognitive science theory, the authors first propose a symbolic-connectionism cognitive architecture. The framework, trying the possibility to unify symbolism and connectionism, which have always been an important branch of AI research [3], is derived from the latest achievements of cognitive science. The purpose of their attempt is to provide a solid foundation for all their research work, and then, under its guidance, research work that the laboratory carried out based on the principles of cognitive science, complex network, and machine learning is illuminated.

The Symbolic-Connectionism Cognitive Architecture

A cognitive architecture is a general proposal about the representations and processes that produce intelligent thought. The most influential cognitive architectures are either rule based, using if-then rules and procedures that operate on them, or connectionist, using artificial neural networks [4]. **The symbolic-connectionism cognitive architecture takes the higher-order predicate-argument structure as the basis of representations and takes systematic similarity computing as the basis of processes** [5].

The predicate-argument structure has always been the primary semantic representation of the sentence in semantics [6]. Representing objects by higher-order predicate-argument structure is one of the theoretical requirements of the structure mapping theory [7]. However, the higher-order predicate-argument structure is redefined by classifying objects into three types: attribute, entity and relation, and those formed phylogenetically from attributes to an entity, from entities to a relation, and from relations to a relation in the systematic similarity model [4]. Such phylogenetic representation brings many benefits: first, following Jeff Hawkins' comments on the characteristics of human intelligence [8], image and language have a unified semantic representation. Second, placing the attribute object in the leaf node of the tree structure of the higher-order predicate-argument structure not only conforms to the

“downgraded predicate” phenomenon that people have already discovered in semantics [6] but also satisfies the priority order of structure mapping [7]. Furthermore, it is even in line with the principle of cognitive economy [4]; that is, the attributes stored in a leaf node can be shared by all its ancestors, which is a storage scheme with the least storage redundancy. There are many other pieces of evidence that also support higher-order predicate-argument structure and similarity computing as the foundation of human cognitive architecture [9, 10]. Synthesizing these new developments in cognitive science, the authors can arrive at the following conclusion: the higher-order predicate-argument structure defined in Ref. [5] is the basic representation of cognitive architecture, and the essence of language understanding is the computation of systematic similarity (and hence is the basic process of cognitive architecture).

In the higher-order predicate-argument structure defined in Ref. [5], attributes, entities, and relations are all concepts. Marked by symbols and referring to objects, concepts are the names of the categories that exist in the human cerebral cortex. By combining multiple argument concepts, a new concept can be formed with the help of a predicate concept or without it. In the former case, the number and semantic roles of arguments are determined by the predicate concept, forming a particular fixed structure, while the latter degenerates into a whole made up of parts. Attribute, entity, and relation may correspond to three different types of neurons that send signals to one another in the cerebral cortex, and a concept is redundantly stored in a certain dendrite or axon of a neuron. In this sense, a network of tens of billions of neurons, which are composed of attribute, entity, and relation neurons, is the representative goal of cognitive architecture [11].

As the primary stage of cognition, the formations of entity concepts, that is, various attributes obtained through the sense organs aggregate to derive the category of a thing, are the processes of perception. After an entity concept forms, the entity combines with other entities and other relations to form a variety of higher-order relations. Since the human learning mechanism is a lifelong continuous learning (regardless of whether one

intends to do so), with the constant growth of human experiences, the processes of combining multiple concepts to form new concepts continue to produce, enrich, and evolve. People learn concepts through analogy, and Bayes’ rule also plays a vital role in concept learning.

The result of learning is the cerebral cortex: a complex network composed of tens of billions of interrelated neurons, which is the massiveness characteristic of the human brain. The relation structures are always updated and changed with the increase of one’s life experience; this is its dynamic nature. It also has concurrency nature, which is confirmed by the fact that people can do multitasks at the same time. Knowledge is always incomplete. Probability is the glue to express the unknown and bond the known.

The symbolic-connectionism representation can give a qualitative interpretation of the effectiveness of deep learning. In the conceptual world, a lower-order structure will always be a component of some higher-order structure so that its value may affect the higher-order structure that contains it. Such systematic characteristic happens to be captured by deep learning with deeper layers of neural network. The systematicness of conceptual space may be the internal reason for the effectiveness of deep learning. Deep learning models the process that multiple attributes form an entity. The result of deep learning training is likely to be infinitely close to a vector that quantifies the attributes necessary to determine an entity.

The knowledge graph is one of the current hot spots in AI research. The most common knowledge in the knowledge graph is head-relation-tail triplet, which is the skeleton of the higher-order predicate-argument structure. Nowadays, the knowledge graph can only be accessed as a static table, so it can only play a minimal role, such as supplementing query information. Under the framework of symbolic-connectionism representation, the tendency of changing from a small static table into massive dynamic concurrent structure drives the development of this technology. Only if the interaction effect mode between a head entity and a tail entity is revealed and a higher-order relation through which the attributes of multipieces of knowledge interact one another

is explored can the potential of the knowledge graph be fully released.

The massive dynamic concurrent probabilistic complex network of conceptual relations is the representation part of the cognitive architecture the authors want to construct, whose complexity is dramatically reduced by the medical field. Combining with the process part, the symbolic-connectionism cognitive architecture can be illustrated using the following concentric circle diagram (Fig. 1).

The authors use the concentric circle diagram to emphasize that the process of outer layer constitutes not only the specific process procedures of its own layer but also the process procedures of all its inner layers. The forming and usage of the representation of symbolic-connectionism architecture lies in the innermost core of the diagram. Systematic similarity computing, as the first usage of the representation, is the basic role of the outer more advanced thinking process. They believe that the research scope of their trying in AIM lies in the core and the first, second, and third layers of the diagram.

Based on the principles of symbolic-connectionism cognitive architecture, treating medicine as the restricted domain, the authors have started a rationalist route towards constructing the virtual brain for clinical reasoning and decision-making, as is fulfilled by (1) defining the knowledge of medicine domain, that is, determining the categorization of entity and entity relation in medicine and researching and realizing IE techniques that are capable of extracting medical entity and medical entity relation automatically from EHRs and other medical data sources; (2) determining medical knowledge representation: the higher-order predicate-argument structure requires head-relation-tail triplet as the basic knowledge representation form; (3) building a medical knowledge base that contains a mass of medical knowledge triplets – researching and realizing complement and expansion methods of the knowledge base based on Bayesian inference, analogy, systematic similarity computing, and lifelong continuous learning methods of continuous knowledge expansion; (4) building reasoning or classification engines based on the knowledge base; and

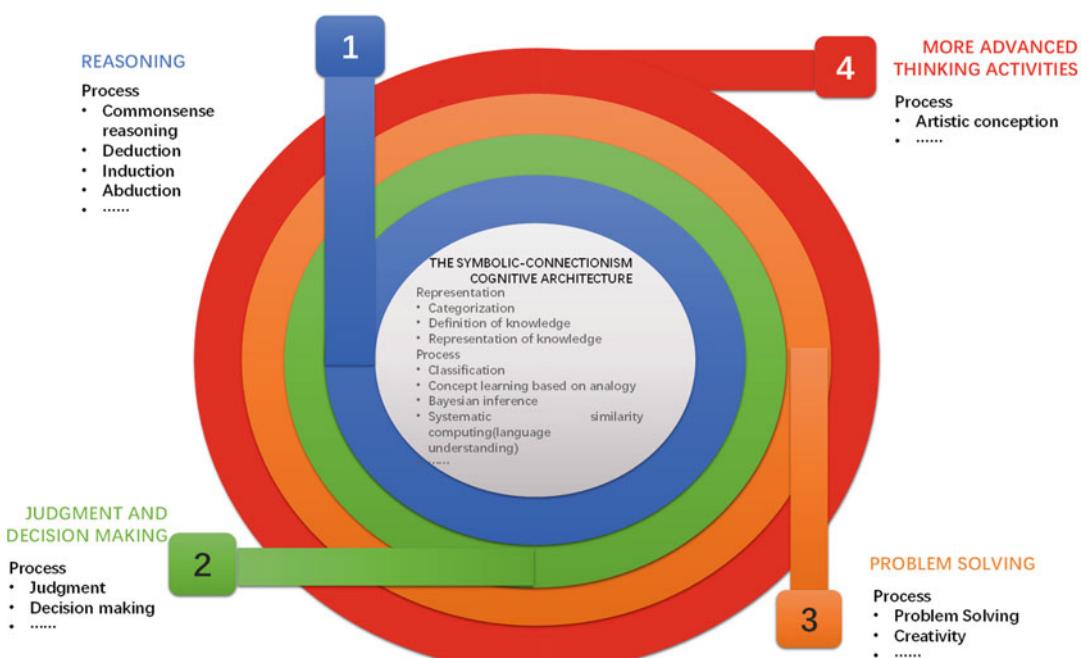


Fig. 1 Hierarchy of thinking

(5) developing applications using these reasoning or classification engines.

The Holistic Framework of Research Work of WI Lab in AIM

Data sources for AIM research include EHRs, medical professional books, medical literatures, medical-related websites, etc. These data sources are collectively called “medical big data” in this chapter. EHR mainly refers to electronic medical record (EMR) in this chapter, that is, the semi-structured digital document that consists of data elements about the medical and treatment history of the patients. EHR also include other structured or semi-structured digital documents prepared by health professionals, such as diabetes nutrition medical record (DNMR) and blood examination records (BERs). The authors use “EHR” and “EMR” interchangeably in the following text. The medical knowledge contained in these data can be divided into empirical knowledge and common sense knowledge according to whether they come from doctors’ medical practice or their education and learning. EMR is a typical data source of empirical knowledge, while the other sources contain both. Clinical text in the unstructured parts of

these data sources shows both generic language features and sublanguage features. Sublanguage features, which are major language features in EMR, denote the language features that are only existent in the unstructured text of the document of some specific types, such as EMR. For example, clinical text in the EMR in Chinese (CEMR) contains such sublanguage features as a number of medical terms, shorthand words and special symbols, templated narrative, occasional ungrammatical expressions, long sentence narrative, etc. Special attention must be paid to deal with these features in NLP or IE tasks.

As shown in Fig. 2, the holistic framework of research and development (R&D) work in AIM can be divided into data tier, knowledge tier, and application tier. Medical entities and entity relations are defined, annotated, categorized, and extracted in the data tier. These basic elements of symbolic-connectionism cognitive architecture are prepared for knowledge tier, where medical knowledge representation and medical ontology construction are implemented. A mass of head-relation-tail triplets, which is the main medical knowledge representation form currently, constitutes medical ontology. These pieces of knowledge are employed by applications in the application tier for classification, reasoning, or decision-making tasks.

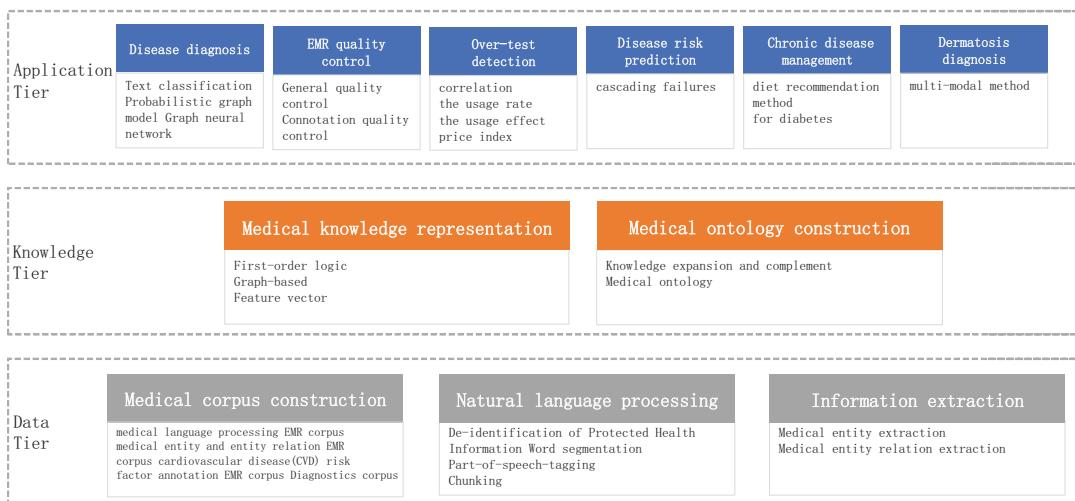


Fig. 2 The holistic framework of research work of WI Lab

Data Tier

Mining the fundamental elements of medical knowledge from medical big data is accomplished in the data tier, as is carried out by three tasks. In the medical corpus construction task, categories of medical entities and medical entity relations must be defined and annotation guidelines and tools must be presented, then medical corpora can be constructed. In NLP task, protected health information (PHI) should be removed firstly, then word segmentation, part-of-speech (POS) tagging, and chunking should be done for the feature extraction of IE. In IE task, medical entities and relations are extracted for knowledge representation. Although IE is usually regarded as a component of NLP, the authors list it separately to highlight its nucleus position in knowledge acquisition.

Four medical corpora, together with their knowledge definitions and annotation guidelines, were constructed by medical professionals (<https://github.com/WILAB-HIT/Resources/>). They are medical language processing EMR corpus [12], for training NLP models for IE; medical entity and entity relation EMR corpus [13, 14], for training IE models for EMR; cardiovascular disease (CVD) risk factor annotation EMR corpus [15], for training CVD risk factor extraction models for EMR; and diagnostics [16] corpus, for training IE models for medical books. The annotation adopts an iterative process to ensure annotative consistency, as shown in stage 2 in Fig. 3.

A risk factor is a pattern of behavior or physical characteristic of a group of individuals that increases the probability of the future occurrence of one or more diseases in that group relative to comparable groups without or with different levels of the behavior or characteristic [17]. Risk factors of CVD, including specific signs such as hypertension and hyperglycemia/diabetes, unhealthy lifestyles such as smoking and alcohol abuse, and other factors such as age and family history, can have prominent effects on the progress of CVD. CVD risk factors have attributes such as indicators, temporal attributes, and assertions that were explicitly or implicitly contained in the records. Besides CVD risk factor itself, these attributions should also be mined. We have built the first annotated corpus of CVD risk factors in

CEMRs, the guideline of CVD risk factor annotations, and the associated IE method to extract them [18]. The risk factor status of a patient can be employed for a personalized modeling of diet recommendation in chronic disease management.

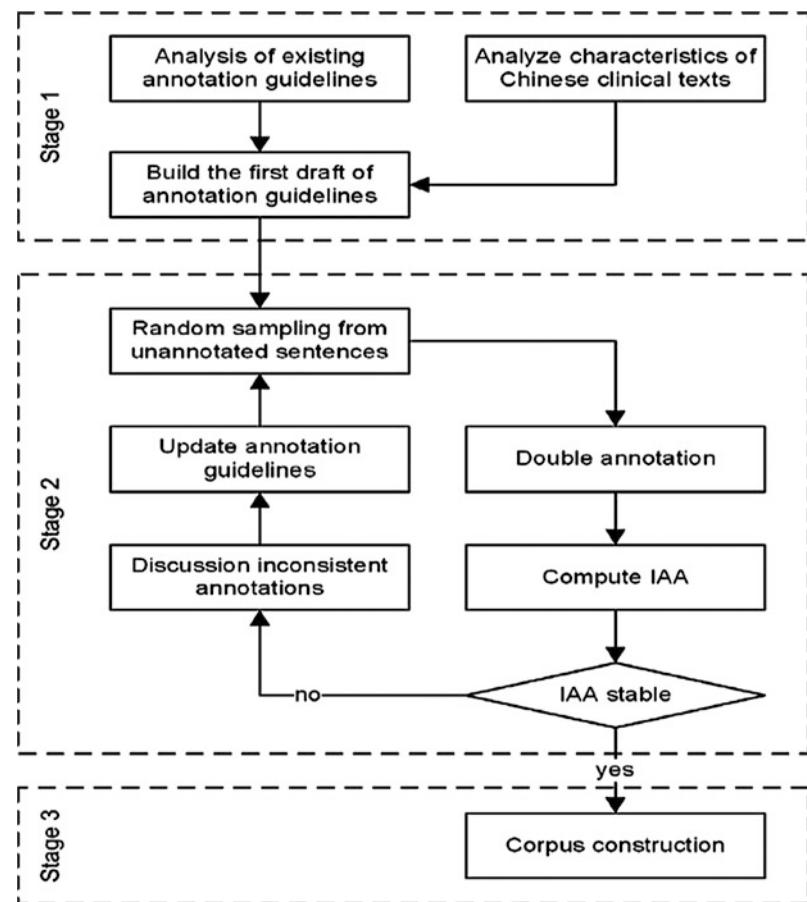
Knowledge Tier

People have tried a variety of methods to define medical knowledge in the diagnosis system. The form of medical knowledge representation is usually inseparable from the diagnosis model. Early attempts include rule based and model based. In recent years, with the development of medical big data, medical knowledge representation has also evolved from the formalization of medical experts' knowledge to automatic acquisition from medical big data. The research of medical knowledge representation is also gradually focused on three aspects: (1) first-order logic based, degenerated from rule-based approaches; (2) graph based, developed from the model-based diagnostic reasoning model; and (3) feature vector based, developed from deep learning. The advantages of first-order logic and its associate reasoning engine are flexibility and powerful expressive ability, but dependence on human labor and low efficiency of reasoning constrain its extensive use. Graph-based representation is a compromise between acquisition difficulty and reasoning efficiency. Feature vector is usually trained by deep learning models and used in classification tasks.

Graph-based representation has been adopted in the researchers' disease diagnosis model in the medical knowledge base that organizes a massive scale of medical triplets. The researchers have also tried the hybrid form of these knowledge representations [19]. And first attempt on the completion and expansion of the medical knowledge base is done [20].

To get close to the symbolic-connectionism cognitive architecture, they have begun their medical ontology construction work, as is carried out, first, by proposing a universal ontology guideline integrating empirical knowledge with common sense knowledge. According to the semantic scheme of the Unified Medical Language System (UMLS) [21] and the annotation guidance of 2010 i2b2/VA

Fig. 3 Iterative annotation method for guideline development and corpus construction. IAA, interannotator agreement



Challenges [22], integrating all the guidelines the researchers put forward before, they design their ontology guideline (<https://github.com/yang1992samantha/MOCG/>). As shown in Fig. 4, they finally define nine kinds of the medical concept – disease, symptom, test, drug, operation, germ, body, department, and health risk factor – and 19 kinds of their relations and attributes. Second is by distinguishing concept and instance. An instance in the medical field refers to a distinguishable and independent medical entity that is in clinical practice. Specifically, entities extracted from the EMRs are called instances. A concept refers to a collection of instances with the same characteristics. The concept of the medical field refers to a collection of summarization and generalization of examples in the medical field with generalization capabilities. The entities extracted from books are called concepts. Concepts are organized in a hierarchical way by a

“subclassOf” relationship, and instance should also belong to the corresponding concept connected by “instanceOf.” Third is by introducing more and more ontological elements into the medical knowledge base. The main feature that distinguishes a medical ontology from a medical knowledge base is the axiom system in which all concepts are defined by formal (description) logic so that any relationship between any two concepts can be derived out by a reasoning engine. The researchers have begun their attempts in this direction. Their current medical ontology contains 82,556 entities and 530,000 pieces of knowledge.

Application Tier

As shown in Fig. 2, the authors have conducted preliminary explorations in six medical applications. Among them, by exploring disease

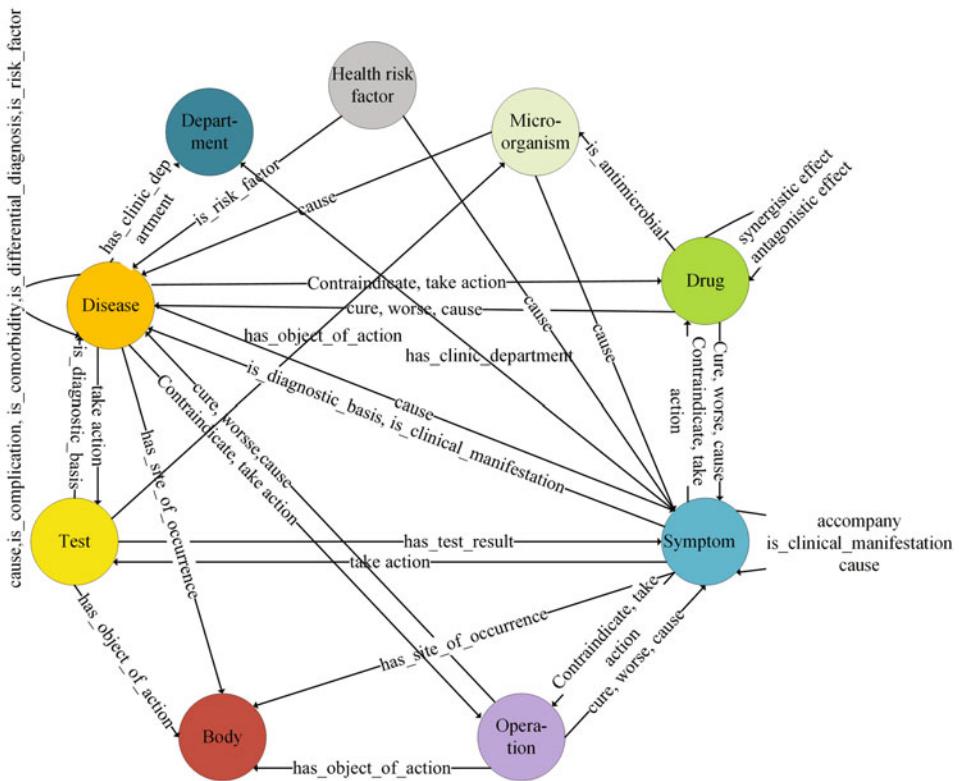


Fig. 4 Medical concepts and their relationship

diagnosis, EMR quality control, disease risk prediction, and dermatosis diagnosis, they study reasoning from different perspectives such as probabilistic reasoning model, application, prediction, and multimodal machine learning, and treat prediction as reasoning on time series data; by exploring overt test detection, they study medical-related decision-making; by exploring diet recommendation methods for diabetes, they study knowledge-based personalized recommendation in chronic disease management.

Disease Diagnosis

Diagnosis is a process in which medical professionals determine the disease from the patient's symptoms or signs. Since clinical reasoning runs through the diagnosis activities, the modeling of clinical reasoning is the most crucial topic of AIM. Nowadays, people's research interests are gradually focused on using classification or reasoning models based on medical knowledge automatically obtained from medical big data.

Classification is one of the most widely researched topics in disease diagnosis. Decision tree, support vector machine, and Bayesian network are designed with a distinct methodology for solving the diagnostic problem, respectively. For improving the accuracy of disease diagnosis, the existing studies have mainly focused on extracting more effective features and exploring more comprehensive pretraining methods while ignoring the importance of domain knowledge.

The probabilistic graphical model is an appealing research direction for reasoning. Many researchers have applied it to disease diagnosis in recent years. In this type of model, the basic knowledge is usually triplets, which can also be regarded as simple first-order logic rules. These pieces of knowledge are then incorporated into a probabilistic reasoning framework such as Markov network, which usually contains a complete theoretical training and probabilistic reasoning process. When these pieces of knowledges are medical knowledge extracted from EMR, a

probabilistic graph model for disease diagnosis is born [23].

The graph neural network is a newly appeared model that shows effectiveness in disease diagnosis. By vectorizing nodes of the knowledge graph, it enriches the interconnectedness between knowledge nodes and provides more ways to reasoning, thus can often bring out more precise disease diagnosis results.

The authors propose an improved probabilistic graph model as a disease diagnosis model and also try graph neural network method to improve the accuracy of disease diagnosis further.

EMR Quality Control

The conventional method of EMR quality control is labor intensive and resource consuming. Using AI technology to raise the effectiveness of quality control has received widespread attention. Beyond some systems, which only pay attention to general quality control like time sensitivity, consistency, and completeness, the researchers' quality control system focus on a more challenging aspect, connotation quality control, that is, the rationality of disease diagnosis.

Under the guidance of “Diagnostics” [16], the researchers summarize three problems to check whether the disease diagnosis in the EMR is reasonable: (1) whether the International Classification of Diseases (ICD-10) codes corresponds to their name – the guidance specifies that each diagnosis must assign its ICD code; (2) whether the principal diagnosis is reasonable – for each EMR, it often contains more than one diagnosis; however, only one or two diagnoses are the reasons for this admission, and they constitute the principal diagnosis; (3) whether the disease diagnosis is complete – each EMR often contains more than one diagnosis. Thus, the researchers need to supplement them according to the content of EMRs.

To solve the above three problems, they construct an EMR quality control system. For each EMR, they first extract all symptom entities and then send them to a disease diagnosis module [23] to achieve a diagnostic list. The diagnostic list contains the top ten diseases and their possibilities. To solve the first problem, they use rules to check whether each disease is assigned to its right ICD code; if not, they return its standard ICD

code. To solve the second problem, they compare the principal diagnosis given by the EMR and the top one diagnosis in the diagnostic list. If they are not consistent, the researchers return their top one diagnosis and let doctors recheck their diagnosis. To solve the last problem, the researchers supplement all diagnosis in the diagnostic list whose possibilities are higher than a predetermined threshold to the EMR diagnosis. Their system now can check the rationality of 20 kinds of respiratory diseases.

Overtesting Detection

Overtesting has become a common phenomenon in the process of reaching a diagnosis, despite the fact that it can cause severe harm to the resources of medical institutions as well as patients' property and health. The authors propose a clinical decision-making framework to determine the appropriateness of diagnostic tests recommended by physicians automatically [24].

Under the guidance of physicians, they first analyze the clinical thinking process and design the unified implicit evaluation criteria to subjectively determine the existence of overttesting for different diseases. And then the implicit evaluation criteria can be modeled by four evaluation strategies, which are the correlation between complaint symptoms and diagnostic tests, the usage rate of diagnostic test t under the condition of related diagnosis, the usage effect of diagnostic test t under the condition of related diagnosis, and the price index. Finally, the four evaluation strategies are carried out through a clinical decision-making framework based on statistical methods integrating a multi-label classifier with a binary classifier.

Disease Risk Prediction

Different from the traditional risk prediction based on time series analysis, the object of disease risk prediction is the occurrence of disease, and disease, especially chronic disease, is often a systemic functional failure rather than a local accidental event. Therefore, the systemic risk prediction method should be used to predict the occurrence of disease, that is, constructing the dynamic evolutionary model of multigranularity and multifactor large-scale complex network to predict the systemic risk of disease.

In progressively serious events, cascading behaviors are considered to be the core driver of risk outbreak, which means failure in one single component would consequently lead to the collapse of the entire system. Inspired by the universal cascading phenomenon, the human body system can also be regarded as a complicated physiological network, in which the interactions between physiological functions maintain the balance of the system. Once one part or function of the physiological network fails, the steady state will be broken, leading to a more significant and more dangerous failure, or even the collapse of the entire network until death, such as complications, infections, immune disorders, etc. To model the physiological domino effect, cascading failures may be a feasible theoretical basis to simulate the deterioration of patients' conditions dynamically.

Chronic Disease Management

Most people will experience two or more chronic diseases during their lives. Nutrition management is an important part of chronic disease self-management. The authors begin their attempt in AI-guided chronic disease management from diet recommendation for diabetes based on nutrition medical records.

For the controlling of CVD risk factors, diabetes is an obstacle that cannot be gotten around. Diabetes, one of the CVD risk factors, is more related to lifestyles, especially diets, than others. The famous Daqing Impaired Glucose Tolerance (IGT) and Diabetes Study proved that positive dietary intervention plays a positive role in the intervention of diabetes and even cardiovascular disease. Diet recommendation is an important way of diet intervention, which can provide patients with a diet plan, closely influence people's daily decisions, and produce long-term significance to health. It is a key technique in the transformation from disease-centered treatment mode to prevention-centered healthcare mode. The authors reviewed the current health recommender system (HRS) research and show that HRSs provide a tool for diet management with health concerns [25].

Diabetes nutrition medical record (DNMR) is a kind of medical record that note documents

medical activities such as inquiry, examination, and treatment of diabetic patients conducted by nutrition doctors. Structurally, DNMR consists of five major components: general status, disease status, laboratory examination, past dietary history, and dietary treatment plan. According to the characteristic of DNMR, as there exists minimum operable quantity (MOQ), which means the minimum quantity of food that can be operated by patients in daily life, the diet recommendation to a patient can be transferred into a multilabel classification problem. Including age, gender, weight, and risk factors, 38 features selected from the DNMR were applied to a random forest-based model. Experiment on 1258 records shows that the method has 0.8448 accuracy of diet recommendation for diabetes. This is the first study that concerns diet recommendation for diabetes based on DNMR. The diet recommendation system has been put into operation in the Nutrition Department of the First Affiliated Hospital of Harbin Medical University. While improving the accuracy of the doctor's treatment, the doctor's treatment time for each patient is reduced from about 15 min to about 5 min (Table 1).

Dermatosis Diagnosis

The difficulties of AIM are not only in research, development, and application; the most challenging point is to become a living AIM application, which means it must become a new positive factor in the ecological environment to which the application field belongs and permanently bring positive effects to all participants in the ecological environment. Taking AI disease diagnosis as an

Table 1 The minimum operable quantity of seven types of food in DNMR

Food type	Minimum value of doctor's advice	Maximum value of doctor's advice	Minimum operable quantity
Staple food	150	375	25
Milk	0	500	50
Egg	0	120	60
Meat, fish, and shrimp	0	100	25
Bean product	0	100	25
Vegetable	50	750	50
Oil	10	35	5

example, a live AI disease diagnosis application must free doctors from tedious daily tasks and increase the reputation and the number of patients of doctors; it must be able to safely and accurately determine patients' diseases and give correct treatment suggestions; it can also bring business opportunities to related industries such as pharmaceuticals and cosmetics and bring benefits and continuous development power to AI application developers. If one of these conditions is not met, it will bring resistance to its own development, which will gradually lose its vitality. Therefore, the application point must be carefully selected to make AI disease diagnosis a living application.

Dermatoses diagnosis is an option with these conditions. Many dermatoses, such as acne, are high-incidence and non-fatal disease. High-precision diagnosis can be achieved only with a photo captured by a mobile phone so that telemedicine on acne is both possible and convenient to accommodate diagnosis AI. At the same time, the multimodal models which combine disease images and texts to make diagnosis is a hot spot of AI researches. The authors have begun their research on acne diagnosis with the support of the department of dermatology of Heilongjiang provincial hospital.

Selected Research

Corpus Construction

To build a comprehensive corpus covering syntactic and semantic annotations of Chinese clinical texts, He et al. [14] propose several annotation guidelines and an annotation method for Chinese clinical texts, providing a foundation for the further study of applying NLP techniques to Chinese texts in the clinical domain. Figure 5 shows the framework of corpus construction on clinical text.

In this work, a comprehensive corpus was built containing annotations of POS tags, syntactic tags, entities, assertions, and relations. Consequently, an IE system was developed based on the annotated corpus. The syntactic corpus consists of 138 Chinese clinical documents with 47,426 tokens and 2612 full parsing trees, while the semantic corpus includes

992 documents that annotated 39,511 entities with their assertions and 7693 relations. IAA evaluation shows that this corpus is of good quality and the system modules are effective. Importantly, it makes a considerable contribution to NLP research into Chinese texts in the medical domain. This corpus is the first comprehensive annotated corpus of Chinese clinical texts, laying a solid foundation for future research.

Information Extraction

He et al. [26] propose a convolutional neural network (CNN)-based architecture with a multi-pooling operation for medical relation classification on clinical records and explore training objective with a relation-type constraint. The model architecture is shown in Fig. 6. Experiments using the 2010 i2b2/VA relation corpus demonstrate these models, which do not depend on any external features, outperform previous single-model methods, and are competitive with the existing ensemble-based method.

He et al. [27] also propose a unified architecture that exploits the advantages of CNN and recurrent neural network (RNN) simultaneously, to identify medical relations in clinical records.

Knowledge Expansion

To enrich and improve the scale of medical knowledge base, Xie et al. [20] propose an incremental expansion framework, which is applied to construct an expandable medical knowledge graph (MKG) in two aspects: (1) integrating external knowledge, which is extracted from the medical information website, and (2) mining potential knowledge, which is hidden in the existing MKG. Figure 7 illustrates the entire expansion framework. First, the authors premise that the integrating procedure and mining procedure are independent. Once a professional or widely recognized external resource is found, the integrating request will be sent to the processing flow. Then a series of knowledge preprocesses are performed, including knowledge extraction from external resource, normalization of medical entities,

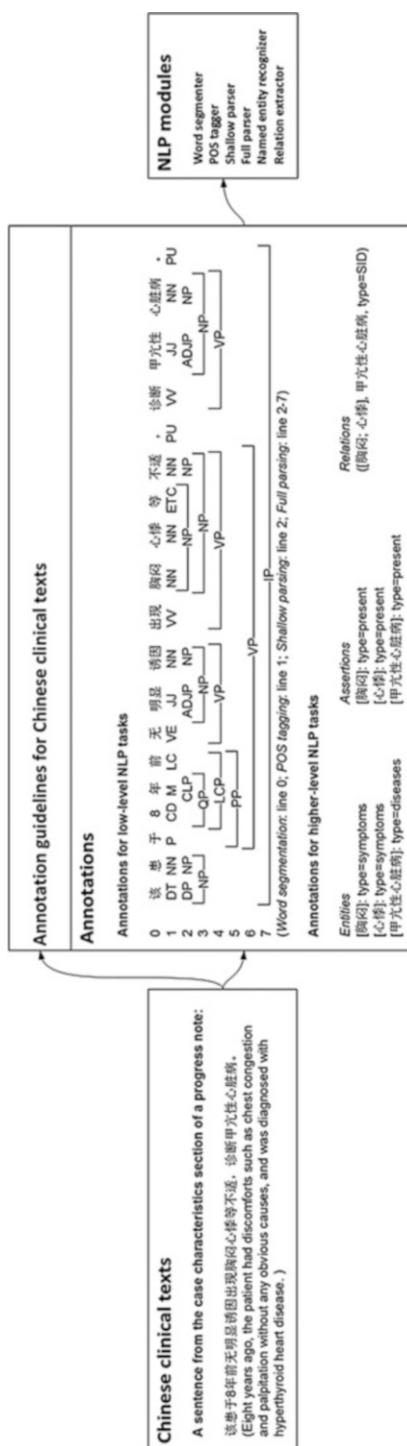


Fig. 5 The framework of corpus construction on clinical text

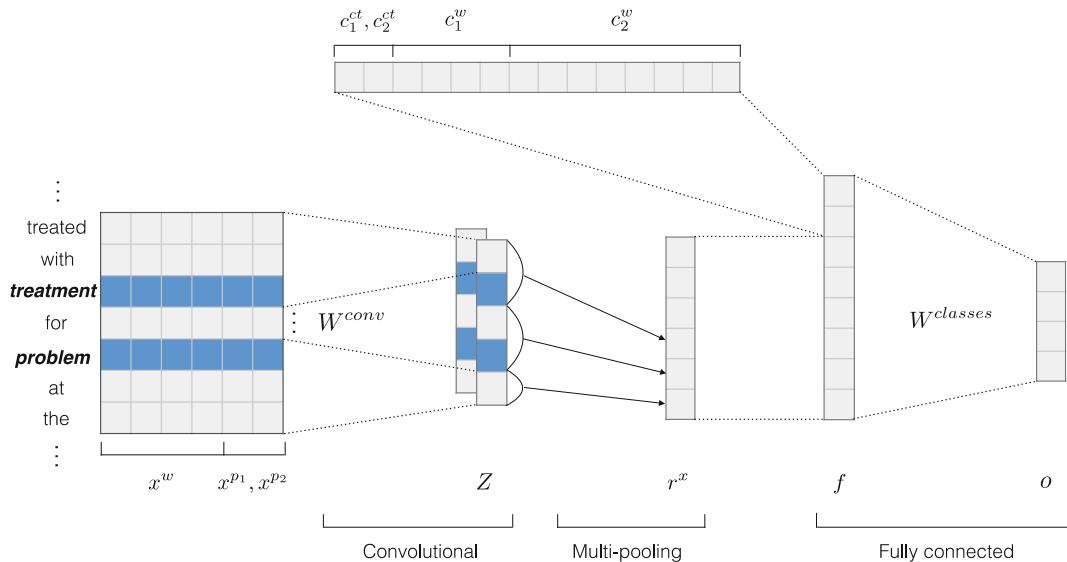


Fig. 6 Architecture of CNN-based model for medical relation classification. Concept contents in the sample sentence “she was treated with [steroids]_{treatment} for [this

swelling]_{medical problem} at the outside hospital, and these were continued” are replaced by their concept types

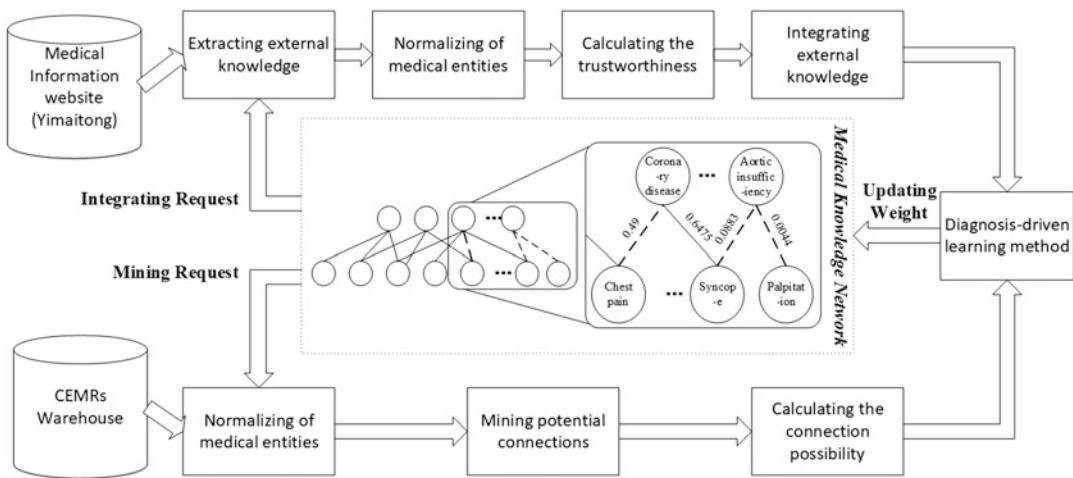


Fig. 7 Workflow of MKN expansion. Solid lines are knowledge from original-MKN, and dotted lines are new expansion knowledge

confidence evaluation of all knowledge, and integration of external knowledge with the Original-MKG.

Integrating External Knowledge

This procedure aims to integrate the disease and symptom entities and their “indication”

relationships with the Original-MKG. When the MKG sends an integrating request, the procedure is performed by four steps: first, by extracting external knowledge. External knowledge is the common sense knowledge that should be distinguished from the empirical knowledge initially. The second step is normalization of medical

entities. Repetitive knowledge will make the network redundant and make diagnostic results biased to the disease which contains more knowledge. Using controlled terminologies (e.g., UMLS) is a regular method to find a medical synonym [28]. All disease entities can be mapped into an ICD-10 code. Considering that most medical synonyms have similar names, the similarity function based on edit distance could be used to normalize these entities [29]. For symptom entities, a symptom vocabulary dictionary is constructed, where symptom concepts with similar semantics share the same code. The third is calculating trustworthiness. To verify the trustworthiness of each candidate's external knowledge, guidelines are collected from the Internet, which have been published over the past ten years and covered most kinds of diseases in different departments. The correlation of the symptom entity S_i and the disease entity D_j is considered as an authenticity index for each external knowledge \mathcal{T}_{ij} . Vector $\mathbf{V}(S_i)$ is constructed to express the state of S_i in clinical guideline set $G = (g_1, g_2, \dots, g_n)$. Disease vector $\mathbf{V}(D_j)$ is formed in the same way. For a candidate knowledge \mathcal{T}_{ij} , the probability $P(S_i, D_j)$ of this knowledge can be calculated by cosine similarity. The fourth step is integrating external knowledge with Original-MKN. The value of $w^*(\mathcal{T}_{ij})$ for knowledge item (S_i, D_j) equals its weight in the Original-MKG, and $w^*(\mathcal{T}_{ij}) = 0$ if \mathcal{T}_{ij} is not contained. Then the external knowledge is fused with the original knowledge to get a candidate weight $w(\mathcal{T}_{ij})$ for each of them by adopting a simple noisy-OR gate:

$$w(\mathcal{T}_{ij}) = 1 - \prod_{k=1}^n \overline{MS_k} = 1 - (1 - P(S_i, D_j))(1 - w^*(\mathcal{T}_{ij})) \quad (1)$$

A knowledge item is considered to be completely impossible only if it had never appeared in any medical resource.

Mining Potential Knowledge

The mining procedure aims to find more potential relationships between existing entities. To evaluate the confidence of each pair of unconnected entities, it is performed by three steps: (1) normalizing the disease and symptom entities, (2) mining potential

connections, and (3) calculating the connection possibility as the candidate weight. The normalization of entities is same as the previous procedure. A detailed description of step 2 and step 3 are given in the following:

1. Mining potential connections: if there is a strong correlation between the neighbors of a disease node and the neighbors of a symptom node, it implies that a potential knowledge exists between two nodes. It is assumed that the symptom might indicate not only the directly linked disease but also the indirectly linked disease, if these two diseases have many similar connected symptoms. For a valid connection $S_i - D_j$, if a path $S_i \xrightarrow{\text{indication}} D_i \xrightarrow{\text{indication}} S_j \dots \xrightarrow{\text{indication}} D_j$ exists, $S_i \xrightarrow{\text{indication}} D_j$ might be connected directly. The indication strength decreases with the path length increasing. Long-distance links indicate weaker indication relationships. Thus, the reasoning path will be confined within a finite-hop neighborhood.
2. Calculating the connection possibility: to obtain the confidence of potential knowledge, the Katz index is used to calculate their connection possibilities [30]. For each potential connection \mathcal{T}_{ij} , the candidate weight is calculated by finding all possible paths between S_i and D_j and summarizing them with damping factor ϵ :

$$w(\mathcal{T}_{ij}) = \sum_{\text{path}_i \in \text{path}} \sum_D \epsilon^d w(\mathcal{T}_{\text{path}_i}^d) \quad (2)$$

where D is the path length and $w(\mathcal{T}_{\text{path}_i}^d)$ is the d -th linking edge weight in the path $_i$. For a potential knowledge, if it has more linking paths and larger weights of linking edges, the indication relationship is stronger.

Clinical Reasoning Model

Knowledge Representation

Some knowledge representation approaches adopt deep neural network to train a language model and obtain indirectly embedding knowledge, of which the entity and relation are

embedded into continuous vector spaces and capture rich syntactic and semantic features from linguistic phrases. Jiang et al. [19] propose a recursive neural knowledge network (RNKN), which transforms medical knowledge based on first-order logic in a hierarchical tree neural network. The model architecture is shown in Fig. 8. Different from some unsupervised embedding methods, RNKN is driven by a disease diagnosis task, and the knowledge vectors are continuously optimized through a large number of EMRs to achieve best performance during disease diagnosis. By reducing the dimensionality of these trained vectors from 150 to 2 using the t-distributed stochastic neighbor embedding (t-SNE) technique, it reveals the interpretation of knowledge embeddings that medical terms with associated semantics are closer together on the scatter plot.

Disease Diagnosis

Zhao et al. [31] give a formal definition of the medical knowledge network (MKN) model. A bipartite graph is constructed by EMRs and denoted by $G = (X, Y)$. If two entities co-occur in the same record, it indicates that there is a specific association between them. Then the weights of all edges are initialized as 0 and depend on modifiers of two entities in EMRs. If their modifiers are both “present,” the weight will be

increased to 1. If the modifier of one of them is “absent,” it indicates a negative correlation, and the weight is decreased to 0.5. The construction of the MKN is finished until all the training records have been added.

Recall that given the status of some entities X , the goal is to infer the status of the unknown entities Y . This problem can be included in a probabilistic inference framework, where the core problem is to calculate the conditional probability $P(Y|X)$. In order to make MKN fit into the inference framework, a Markov random field (MRF) is applied to treat nodes as random variables. These variables can take real values from 0 to 1 and indicate the degree of these entities on the given patient. The entities with positive assertion are denoted as 1, while those that do not appear or with negative assertion is set to 0. The core part of MRF is to define an energy function f (y, x) and model the signs of each $x \in X$ and each $y \in Y$. When $f(y, x) < 0$, the model prefers that y and x take the same sign, which means that a patient with status x as positive are likely to have status y as positive. Otherwise, it implies that y and x are more likely to have different signs.

There are two different strategies to define $f(y, x)$. The first is to assign a real-valued learnable parameter to each (y, x) pair. The second is to assign a dense vector representation for each nod, and

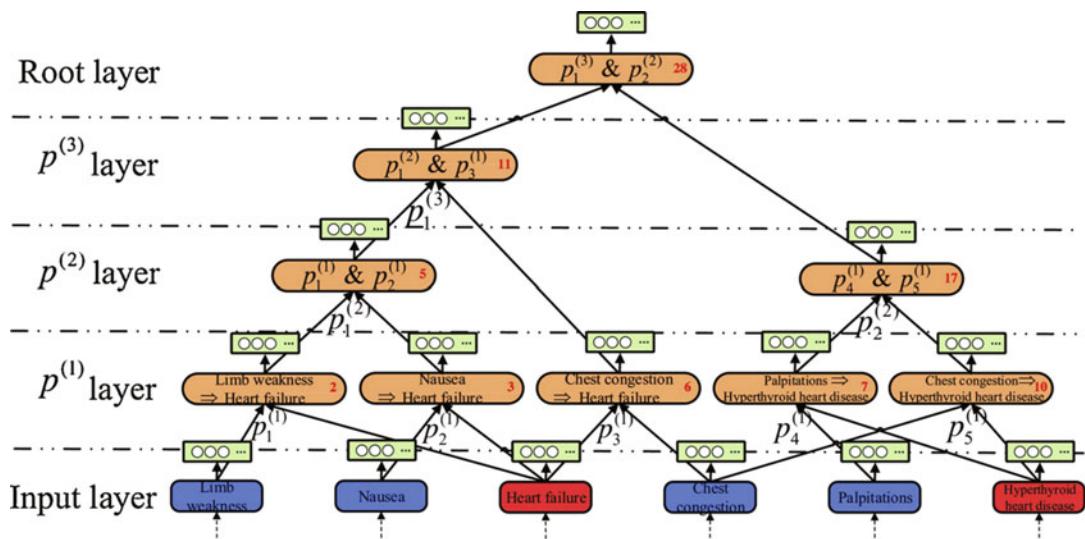


Fig. 8 Huffman tree of the RNKN, representing deep first-order logic knowledge for multidisease diagnosis

calculate the energy function of (y, x) by bilinear transformation or translating distance. Notice that in the first strategy, the parameters are assigned to edges, but in the second strategy node representation is parameterized, which reduces the parameter set from $\Theta(|Y| \times |X|)$ to $\Theta(|Y| + |X|)$. An entire framework is built to learn the parameters from the data and make the inference accordingly.

Jiang et al. [23] propose another multivariate inference methodology that combines a weighted knowledge graph with probabilistic models. By incorporating a Boltzmann machine into the potential function of the Markov network, a joint probability distribution of clinical variables is defined as the following expression:

$$P(X = x) = \frac{1}{Z} \exp \left(\sum_{\substack{i \\ r_i \in R}} \omega_i x_{r_i}^s x_{r_i}^d + \sum_j \left(\sum_{z=1}^n m_z e^{-\sigma} \right) \cdot x_j \right) \quad (3)$$

This joint probability distribution rewrites the potential function of Markov network $\phi(r) = \exp(-\varepsilon(r))$ with a parametric energy function $\varepsilon(r) = \varepsilon(x_d, x_s) = -\omega_{d,s} x_d x_s$, where the medical triplet r consists of a disease entity d and a symptom entity s and x reflects the severity of each entity. In addition, the Gaussian potential function (GPF) is employed to evaluate the potential energy of each node, which is expressed as

$$\varepsilon(x_j) = u_j x_j = \left(\sum_{z=1}^n m_z e^{-\sigma} \right) \cdot x_j \quad (4)$$

where σ represents the distance coefficient between entity j and its neighbor z in knowledge graph and m_z is the quality of neighbor z . According to the definition of joint distribution, two common generic clinical questions are answered in this work: “What is the probability that rule r_1 holds given rule r_2 ?” and “What is the probability of disease d_1 given the symptom vector S ?” In a multidisease

diagnosis task, the experimental results on the actual EMRs and BERs demonstrate that the proposed inference framework significantly outperforms the general machine learning baseline methods such as logistic regression and neural networks. It further proves the effectiveness and promise of the novel knowledge-driven clinical reasoning.

Disease Risk Prediction

The first issue is how to establish an expressive physiological network. In clinical medicine, the abnormalities in vital signs are the internal cause of disease. Meanwhile, disease is the external manifestation of abnormal signs. When a disease emerges, it indicates that some related signs are falling in borderline or abnormal states. Sequentially, if the severity of these abnormal signs exceeds tolerance thresholds of other diseases, it will raise the failure risk of more physiological functions, even cause a chain reaction. Here, the physiological system is formalized as a bipartite network $\mathcal{G} = (\varepsilon_e, \varepsilon_i, \mathcal{R}, \mathcal{W})$, where ε_e and ε_i are the set of disease nodes in subnetwork \mathcal{G}_e and the set of sign nodes (e.g., laboratory examination items) in subnetwork \mathcal{G}_i , respectively. \mathcal{R} is the set of interaction edges between two subnetworks, and \mathcal{W} represents the weight set of triplets. For generality, network X means either subnetwork \mathcal{G}_e or \mathcal{G}_i . If the number of surviving neighbor nodes $k_{s,i}$ that are connected to node i by interaction edges remains greater or equal to tolerance threshold k_i^* , node i with k_i -degree remains functional and transforms its state S_i to 1; otherwise, $S_i = 0$. The tolerance coefficient $r_i(k_{s,i}, k_i)$ can be a step function as

$$r_i(k_{s,i}, k_i) = \begin{cases} 0 & k_{s,i} < k_i^* \\ 1 & k_{s,i} \geq k_i^* \end{cases} \quad (5)$$

a linear function as $r_i(k_{s,i}, k_i) = k_{s,i}/k_i$, or a nonlinear polynomial function as $r_i(k_{s,i}, k_i) = 3(k_{s,i}/k_i)^2 - 2(k_{s,i}/k_i)^3$, which has been used in Ref. [32].

The second issue is how to depict the risk propagation caused by initial failure nodes.

When either of the two connected nodes fails, their interacting edge is also removed; e.g., the emergence of renal failure disease will cause the kidney to lose the physiological function of detoxification. As a result, there is no longer a supply interaction to maintain the physical stability of urine osmotic pressure and plasma albumin. Referring to the cascade analysis of complex network in theoretical physics [33], the generating functions of degree distribution and the excess degree distribution for two subnetworks can be used to measure the robustness of the physiological network at each cascade stage. The generating function of degree distribution [34] in X network is defined as follows:

$$G_X^0(q) = \sum_{k=0}^{\infty} P_X(k) \sum_{k_s}^k \binom{k}{k_s} q^{k_s} (1-q)^{k-k_s} r(k_s, k) \quad (6)$$

where $P_X(k)$ is the degree distribution of X network, q is the probability of a randomly chosen edge, which is connected to a functional node. The generating function of the excess degree [35] is an extension of $G_X^0(q)$ with respect to functional edges:

$$G_X^1(q) = \sum_{k=0}^{\infty} \frac{kP_X(k)}{k_X} \sum_{k_s}^{k-1} \binom{k-1}{k_s} q^{k_s} (1-q)^{k-k_s-1} r(k_s + 1, k) \quad (7)$$

where k_X is the average degree of network X.

Proof Given a randomly connected network, the generating function of the degree distribution $P(k)$ is defined as

$$G_X^0(q) \equiv \sum_{k=0}^{\infty} P_X(k) x^k \quad (8)$$

It follows from Eq. (9) that the average degree of the network is $k = \sum_{k=0}^{\infty} kP(k) = G'_0(x) \Big|_{x=1}$. Following a randomly chosen edge, the probability of reaching a node with k outgoing edges (the degree of the node is $k+1$) is

$$\tilde{P}(k) = \frac{(k+1)P(k+1)}{\sum_{k=0}^{\infty} [(k+1)P(k+1)]} \quad (9)$$

Notice that

$$\begin{aligned} & \sum_{k=0}^{\infty} (k+1)P(k+1)x^k \\ &= \sum_{k=0}^{\infty} kP(k+1)x^{k-1} = G'_0(x) \end{aligned} \quad (10)$$

and

$$\sum_{k=0}^{\infty} [(k+1)P(k+1)] = G'_0(x) \Big|_{x=1} = \langle k \rangle \quad (11)$$

The generating function for the distribution of outgoing edges $\tilde{P}(k)$ is

$$\begin{aligned} G_1(x) &\equiv \sum_{k=0}^{\infty} \tilde{P}(k)x^k = \frac{G'_0(x)}{k} \\ &= \sum_{k=0}^{\infty} \frac{kP(k)}{k} x^{k-1} \end{aligned} \quad (12)$$

Then the given random network (with degree distribution $P(k)$) is replaced by a tree structure. Each of these nodes is in turn connected to $k-1$ neighbors at the next lower level. The levels of the tree are labeled from $n = 0$ at the bottom, with the top node at an infinitely high level ($n \rightarrow \infty$). Define q_n as the probability that a node on level n is functional. Consider updating a node on level $n+1$, assuming that the nodes on all lower levels have already been updated. With probability $\tilde{P}(k)$, the chosen node has k neighbors: one of these is its parent (on level $n+2$), and the remaining $k-1$ are its children (on level n). Each of the $k-1$ children is active with probability q_n . Thus, the node has m active children (and therefore $k-1-m$ inactive children) with probability $\binom{k-1}{m} q_n^m (1-q_n)^{k-1-m}$. The probability of distribution of its tolerance coefficients is given by function $r(m+1, k)$, and combining the independent probabilities yields the equation for q_{n+1} :

$$\begin{aligned}
q_{n+1} &= G_1(q_n) \\
&= \sum_{k=0}^{\infty} \frac{kP(k)}{k} \sum_{m=0}^{k-1} r(m+1, k) \binom{k-1}{m} q_n^m \\
&\quad (1 - q_n)^{k-1-m}
\end{aligned} \tag{13}$$

The final issue is how to model a general process of cascading failure in the medical field. If a fraction $1 - y_{G_e,1}$ of nodes is removed from ε_e to indicate the emergence of disease, the fraction of surviving nodes $y_{G_e,1}$ in G_e will determine the connectivity of the current bipartite network. Considering that another network G_i has not been affected by the first cascade, the functionality of a randomly chosen interaction edge $q_{G_e,1}$ only depends on $y_{G_e,1}$. Thus, the equation $q_{G_i,1} = y_{G_e,1}$ is set up. Applying generating function $G_X^0(q)$ to G_i , the fraction of surviving nodes in G_i will be updated by $y_{G_i,1} = G_{G_i}^0(q_{G_i,1})$. Because network G_e has been damaged at the second cascade stage, $q_{G_e,2} \neq y_{G_i,1}$. The excess degree distribution is used to calculate $q_{G_e,2}$, denoted as $q_{G_e,2} = G_{G_i}^1(q_{G_i,1})$. Through inducting the general cascading mechanism, the recursion relations for stages $l > 1$ can be formalized:

$$q_{G_e,l} = G_{G_i}^1(q_{G_i,l-1}); q_{G_i,l} = y_{G_e,l} G_{G_e}^1(q_{G_e,l}) \tag{14}$$

The fractions of surviving nodes of two sub-networks at l -th stage are

$$y_{G_e,l} = y_{G_e,1} G_{G_e}^0(q_{G_e,l}); y_{G_i,l} = G_{G_i}^0(q_{G_i,l}) \tag{15}$$

In conclusion, the robustness of a physiological network and the characteristics of failed nodes are estimated iteratively by Eq. (14) and Eq. (15) at each cascade stage. If these temporal evolution features are fed into sequential models, such as long-short term memory (LSTM), gated recurrent unit (GRU), and Transformer, the deterioration of patients' condition will be retraced and the health risks can be continuously predicted.

Conclusion and Future Work

Based on EHRs and other medical data sources, the authors have successfully carried out a series of AIM research and development, and some results have been put into actual operation. However, their research is still in an initial state, as is mainly reflected in the following aspects: in knowledge acquisition, the authors have tried medical entity and relation extraction based on deep learning methods, among which entity relation extraction is still a bottleneck that impedes the overall precision. In knowledge representation, triplets are their primary way, but this way has increasingly shown low expressiveness and is not suitable for rich and complex medical knowledge. They need to adopt a more expressive form of knowledge that can be automatically obtained from massive medical big data. In medical ontology building, the total amount of medical knowledge is still relatively small; the axiom system that gives the ability to reason between concepts is under construction. Knowledge complement and expansion method are still in their infancy stage, not to mention concept learning based on analogy and systematic similarity, knowledge complement based on Bayesian inference, and lifelong continuous learning mechanism. In terms of reasoning, the probabilistic reasoning model they proposed can only produce beneficial results from a small amount of knowledge and needs to be further improved in both ability and accuracy. The key to solving these problems is to break through the bottleneck of knowledge representation. From the discussion of cognitive architecture, the authors can understand that the massiveness, dynamics, concurrency, and interconnectivity of the human cognitive architecture are unmatched by the triplet. Therefore, how to overcome the limitation of triplets and study a new medical knowledge representation with characteristics closer and closer to the human brain is the direction of their future efforts. This requires them to propose an entirely new type of medical knowledge representation with innovation in terms of the system architecture.

Acknowledgment This research was financially supported by the Open Research Fund from Shenzhen Research Institute of Big Data, Shenzhen 518000, under Grant No. 2019ORF01011.

References

1. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med.* 2019;25:24–9. <https://doi.org/10.1038/s41591-018-0316-z>.
2. Patel VL, Arocha JF, Zhang J. Thinking and reasoning in medicine. In: Holyoak KJ, Morrison RG, editors. *The Cambridge handbook of thinking and reasoning*. Cambridge University Press; 2005. p. 727–50.
3. Sun R, Alexandre F. Connectionist-symbolic integration: from unified to hybrid approaches. L. Erlbaum Associates; 1997.
4. Thagard P. Cognitive architectures. In: Frankish K, Ramsey W, editors. *The Cambridge handbook of cognitive science*. Cambridge University Press; 2012. p. 50–70.
5. Guan Y, Wang X, Wang Q. A new measurement of systematic similarity. *IEEE Trans Syst Man Cybern A Syst Humans.* 2008;38(4):743–58.
6. Leech G (1981) Semantics, penguin books (New York, NY), 1974, revised edition
7. Gentner D. Structure mapping: a theoretical framework for analogy. *Cogn Sci.* 1983;7:155–70.
8. Hawkins J. On intelligence. 1st ed. Times Books; 2004. p. 272.
9. James R. Hurford, the neural basis of predicate-argument structure. *Behav Brain Sci.* 2003;26:261–316.
10. Deyi L. Formalization of brain cognition: design discussion of driving brain. *Sci Technol Rev.* 2015; 33(24):125.
11. Carter R. *The human brain book: an illustrated guide to its structure, function, and disorders*. DK Publishing; 2009.
12. Jiang Z, Zhao F, Guan Y (2014) Developing a linguistically annotated corpus of Chinese electronic medical record. In: IEEE international conference on bioinformatics and biomedicine (BIBM), pp 307–310
13. Yang JF, Guan Y, He B, Qu CY, Yu QB, Liu YX, Zhao YJ. Corpus construction for named entities and entity relations on Chinese electronic medical records. *Ruan Jian Xue Bao/J Softw.* 2016;27(11):2725–46. (in Chinese). <http://www.jos.org.cn/1000-9825/4880.htm>
14. He B, Dong B, Guan Y, Yang J, Jiang Z, Yu Q, Cheng J, Chunyan Q. Building a comprehensive syntactic and semantic corpus of Chinese clinical texts. *J Biomed Inform.* 2017;69:203–17.
15. Jia S, He B, Guan Y, Jiang J, Yang J. Developing a cardiovascular disease risk factor annotated corpus of Chinese electronic medical records. *BMC Med Inform Decis Mak.* 2017;17(1):117_1–117_11.
16. Wan X, Xuefeng L. *Diagnostics*. 8th ed. People's Medical Publishing House; 2013.
17. Rothstein WG. Public health and the risk factor: a history of an uneven medical revolution. Rochester: University of Rochester Press; 2003.
18. Jia S, Hu J, Jiang J, Xie J, Yang Y, He B, Yang J, Guan Y. Extraction of risk factors for cardiovascular diseases from Chinese electronic medical records. *Comput Methods Prog Biomed.* 2019;172:1–10.
19. Jiang J, Wang H, Xie J, et al. Medical knowledge embedding based on recursive neural network for multi-disease diagnosis. *Artif Intell Med.* 2020;103: 101772.
20. Xie J, Jiang J, Wang Y, Guan Y, Guo X. Learning an expandable EMR-based medical knowledge network to enhance clinical diagnosis. *Artif Intell Med.* 2020; 107:101927.
21. Olivier B. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 2004;32:267–70.
22. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc.* 2011;18:552–6.
23. Jiang J, Li X, Zhao C, et al. Learning and inference in knowledge-based probabilistic model for medical diagnosis. *Knowl-Based Syst.* 2017;138:58–68.
24. Yang Y et al. A clinical decision-making framework against over-testing based on modelling implicit evaluation criteria. <https://github.com/yang1992samantha/CDF-MIECO>
25. Su J, Guan Y, Li Y et al (2020) Do recommender systems function in the health domain: a system review. arXiv preprint arXiv:2007.13058
26. He B, Guan Y, Dai R. Classifying medical relations in clinical text via convolutional neural networks. *Artif Intell Med.* 2019;93:43–9.
27. He B, Guan Y, Dai R. Convolutional gated recurrent units for medical relation classification. In: 2018 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE; 2018. p. 646–50.
28. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of the AMIA symposium. American Medical Informatics Association; 2001. p. 17.
29. Navarro G. A guided tour to approximate string matching. *ACM Comput Surv (CSUR).* 2001;33(1): 31–88.
30. Lü L, Zhou T. Link prediction in complex networks: a survey. *Physica A Stat Mech Appl.* 2011;390(6): 1150–70.
31. Zhao C, Jiang J, Xu Z, et al. A study of EMR-based medical knowledge network and its applications. *Comput Methods Prog Biomed.* 2017;143:13–23.
32. Di Muro MA, Valdez LD, Rêgo HHA, et al. Cascading failures in interdependent networks with multiple supply-demand links and functionality thresholds. *Sci Rep.* 2017;7(1):1–10.
33. Newman MEJ, Strogatz SH, Watts DJ. Random graphs with arbitrary degree distributions and their applications. *Phys Rev E.* 2001;64(2):026118.
34. Gleeson JP, Cahalane DJ. An analytical approach to cascades on random networks. Noise and stochastics in complex systems and finance. *Int Soc Optics Photon.* 2007;6601:66010W.
35. Shao J, Buldyrev SV, Braunstein LA, et al. Structure of shells in complex networks. *Phys Rev E.* 2009;80(3): 036105.



AIM and Causality for Precision and Value-Based Healthcare

19

Hector Zenil

Contents

Current Standard of Care and Medical Research	287
Current AI Practice in Medicine and Healthcare	288
Model-Driven AI and Causal Diagnostics	290
The Opportunity for Precision Medicine	290
Conclusions	291
References	291

Abstract

AI has the potential to transform medicine and healthcare, but the value of current applications, often exciting and sometimes helpful, has been overstated. The responsible application of AI, however, especially in advancing medicine as a scientific endeavor, has great potential, which is only just beginning to be tapped. We propose using AI to move from population-wide symptom pattern matching toward individual causal diagnostics, and to advance healthcare from only cost-effective

and population-wide pattern matching toward patient-centered, value-based, predictive, and precision care.

Keywords

Causal diagnostics · Precision medicine · Value-based healthcare · Personalized medicine · Model-driven AI · Causation · Diagnosis and prognosis

Current Standard of Care and Medical Research

Laboratory tests cost the UK's National Health Service (NHS) billions per year, and according to a recent report, this use of resources by the NHS could be curtailed. For example, up to 80% of the total number of tests are only requested in tandem because of long turnarounds, making doctors reluctant to wait for the result of a first test

H. Zenil (✉)
Oxford Immune Algorithmics Ltd, Reading, UK
Alan Turing Institute, London, UK
Algorithmic Dynamics Lab, Unit of Computational Medicine, Karolinska Institute, Stockholm, Sweden
e-mail: hector.zenil@algocyte.ai; hzenil@turing.ac.uk; hector.zenil@ki.se

before ordering a possible second one, which may turn out not to be necessary in certain cases. Meanwhile, requests for blood test are rising in number as they are the first step in any initial health screening, and this trend is expected to continue during the pandemic.

A General Practitioner spends around 2–3 h a week on average checking patients' blood results, and almost 40,000 additional GPs will be needed in the next 8 years. Thanks to Covid 19 and the NHS' Long-Term Plan, online appointments have been increasing in number. However, many of these patients must still travel to hospital for things like blood tests. In pediatric cancer care, for example, 9,500 children undergo regular full blood count tests once per week, which places a heavy financial and mental burden on these children and their parents. What's more, during the pandemic these children haven't been properly monitored, being vulnerable patients with compromised immune systems.

The current approach to preventing people from having to visit hospitals consists of sending a nurse into the community, which entails significant overhead costs that make it anything but cost effective. On the one hand, not only is the cost of such visits 20 times higher than the cost of, e.g., a blood test, but staff time is taken away from other tasks.

Current approaches also pose obstacles to changing the way blood testing is done because they are based on very expensive equipment that requires specialized operators and comes with very onerous annual maintenance contracts which are usually as or more expensive than the machinery itself. Current technology, which consists of very large, expensive machines that are not user-friendly, are unable to allow the regular testing needed for early diagnosis and proper health monitoring.

Moreover, biology and medicine are heavily multiscale making it difficult to model at any single level [2]. Instead, powerful model-driven approaches are needed in the study of complex diseases [7] to overcome the curse of data dimensionality, with a model able to extract main features and selection to be able to abstract the key elements of a disease diagnosis and prognosis.

However, scientists and clinicians alike are unable to keep abreast of their own areas of expertise, with individual scientists increasingly ignorant of advances not only in all other scientists' fields but even in their own. For example, only between 2003 and 2016 alone, the total number of scientific papers published annually more than doubled [1] and the doubling period is being shortened by about half every 5 years. Things are set to get worse, not better, with today's output at almost 2.5 million papers per year. Only AI could help close this gap by digesting, selecting, and exploiting this information faster.

Eventually, only AI will be able to catch up, leaving human scientists to guide AI on the type of meaningful research that should be conducted, with a focus on what scientists believe is of greatest interest. Most of the work will then be a matter of translating back and forth between humans and AI, providing instructions to AI and interpreting results from AI so humans could understand them.

Current AI Practice in Medicine and Healthcare

While techniques such as machine and deep learning (e.g., deep neural networks) are very effective in performing tasks related to classification, they are only so under highly controlled conditions, and are ill-equipped to deal with basic cognitive functions that we as humans take for granted, including basic logical inference and abstraction for model generation and hypothesis testing.

The combination of symbolic computation and statistical machine learning is thus a powerful combination for tackling complex problems, and a way forward to achieve better and more responsible A.I. This is because it offers interpretable models that correspond to observations, and can be tested, rather than encoded in obscure black-box explanations.

Driven by digitalization and rapid progress in the field of artificial intelligence (AI), and particularly in the field of machine learning (ML), the healthcare industry is undergoing radical changes that are constantly accelerating. The potential for

improvements in healthcare is significant if done properly. The responsible application of AI procedures can derive findings and recommendations for action from the vast amounts of data constantly being produced in digital form (Big Data). This is particularly relevant in cancer research, where new non-invasive imaging technologies are opening up completely new fields of research, and where existing imaging modalities like magnetic resonance imaging (MRI), computed tomography (CT), and positron emission tomography (PET) are extensively used over the complete cycle of patient care, from the early detection and diagnosis of tumors to the evaluation of treatment responses and follow-up.

One overarching reason for the small number of AI applications that make it to clinical application in general, is the lack of trust in AI-based applications, which are perceived as black boxes producing results on the basis of non-transparent decision making, and produce failures that are difficult to understand and unpredictable. This meta-problem is rooted in multiple individual but strongly related issues with AI development and validation.

Big data-based AI technologies hold still potential but model-driven AI which is explainable and can contribute to medical understanding, can better contribute to value-based healthcare for better outcomes through better prediction, diagnosis, and treatment of diseases.

There is also a lack of methods and protocols for the development, testing, and validation of reliable and reproducible AI-based medical approaches without the traditional hype. There is a non-trivial issue of measuring the degree of AI involvement and its relevance and impact. While quality is commonly assessed using performance measurements, a comprehensive evaluation of the quality of AI-based solutions should consider various other criteria, including robustness, explainability, and actual impact on clinical decisions.

These issues within the current AI development and validation cycle represent enormous obstacles to the practical use of AI outside shallow applications, that is, applications that do not offer scientific insight and are chiefly related to categorization tasks.

To illustrate the limitations of statistical AI, no amount of big data or statistical massaging (no matter how ‘sophisticated’) can make a traditional deep neural network perform a simple arithmetical operation if it has not seen it before, because it has no concept of number or symbol and will treat the whole term as an object.

One of the main problems data scientists face when working with images or medical data in the medical domain is the typically small amount of publicly available data, data being crucial for training and validation of neural networks. Such obstacles are especially challenging to overcome for start-ups and small commercial enterprises (SMEs) lacking the required resources and contacts for scaling their projects and accessing the first healthcare market.

The question is thus how AI can partner with clinicians and patients to improve quality of care. If we want AI and humans to collaborate, it is very important to recognize their respective capabilities and skills. In fact this collaboration began a long time ago, and it was accompanied by certain implicit understandings. For example, when calculators, cash registers, and electric counting machines entered the market, it was because it was acknowledged that these machines performed the tasks they were designed for far better and faster than their human counterparts. With AI as a newly developed computational asset, a similar process is afoot, for example, in the recognition that deep neural networks appear to be in some ways better at categorizing objects than humans. Hence the hype regarding new AI systems able to characterize and categories all sorts of medical images and data often purportedly better than clinicians, and if not always better, then faster. However, some subtleties are worth noting. While deep learning classifiers are faster once trained, they are also several orders of magnitude slower at learning or training, require much more data, and more importantly, are currently far less robust than their human counterparts. Humans build models based on function or highly abstracted features. Take, for instance, a school bus. It is its purpose – transporting children to and from school – that is the foremost consideration in its design, and only secondarily its looks, size,

shape, and color. A recognition of capabilities that evolve over time is key to optimally partnering with AI, even as AI improves in areas where its abilities are currently weak.

Model-Driven AI and Causal Diagnostics

One of our main advantages is how we approach AI by combining the best of different approaches: neural networks, that have produced incredible results in classification tasks, and what is called symbolic computation. The best way to understand symbolic computation is to think of a traditional calculator that understands the concept of number to the level of being able to perform arithmetic operations and follow an algorithm, which is something that neural networks find difficult to accomplish. We think this approach is a responsible way to control some variables that are key to being able to better understand and explain some of the decisions that an AI algorithm may take in medicine and health care.

An area that has also seen advances and great interest is explainable AI systems, that is, AI able to explain in human terms what, why and even how they do what they do or go about making a decision.

However, despite the many advances in medical technology, practical clinical diagnosis is still largely based on statistical symptom pattern-matching. The idea that a cluster of symptoms can lead to the correct diagnosis is often an oversimplification. Not only are symptoms not binary (either one has it or not), but complex diseases are multifactorial and can display a wide number of dissimilar effects and confounding signals. To better deal with probable root causes, we propose leveraging the results of years of research in model-driven causal discovery [3–5] by developing the area of causal diagnostics, a cutting-edge and responsible model-based approach to medicine different from simplistic AI approaches based on statistical classification that are typically unable to provide generative models or explanations for their arbitrary choices.

Longitudinal, non-invasive, (quasi) real-time, and cost effective health monitoring is hugely

important if we are to improve public health and deliver on the promise of personalized medicine for timely, accurate diagnosis.

The training and inference engines behind causal diagnosis are based on what is identified as neuro-symbolic computation, a combination of the best AI tools with cutting-edge expert systems that rely on both mind and machine to fight disease.

Causal diagnostics can revolutionize medicine by refocusing on first principles and causal generating mechanisms to redesign treatment and drugs so that they go beyond offering symptom relief. We call our neuro-symbolic computation approach a type of responsible Artificial Intelligence (AI) because it goes beyond current black-box approaches in AI in order to reach a deeper understanding of critical questions relevant to the unveiling of root causes in areas of medicine.

The Opportunity for Precision Medicine

Watson, for example, was at the pinnacle of statistical inference. Its success on the TV show Jeopardy! was never matched by successes in industry, in particular in areas of medicine and healthcare, where IBM initially thought it could be implemented and could make a real difference. Watson may have led to some marketing strategies in IBM's cognitive computation campaign but the methods behind the success of Watson at Jeopardy! were difficult to translate into other areas.

The problem that Watson and other AI approaches have tried to solve is, in our opinion, at once easier and more difficult.

First, the whole area of automation and AI application to medicine and healthcare is underdeveloped, and bringing to bear so much computational power at such a high cost, and overpromising while underdelivering, seems to have been the wrong approach. Instead, we believe that starting simpler is better, and then moving step-by-step, adopting only processes that are already known to work in the field and improving them with AI can prove to be a more robust approach after exhaustive evaluation.

In an application to risk assessment of blood test results, for example, we have developed an innovative score used in our health diagnostics and analytics methods. The score was tested against 100,000 cases, demonstrating that it picks up signals to separate healthy from unhealthy groups, even from noisy sources (the NHANES database from the CDC).

The score introduced in [6] is a numerical score that goes from 0 (healthy) to 10 (unhealthy), quantifying how removed a person's individual values are from reference population values, and is able to separate healthy from unhealthy groups even in the face of noise.

In a further improved version of these risk assessment scores, the scores adapt and learn over time with incremental precision each patient's baseline, allowing the application to flag for deviations smaller than the traditional reference population values, and thereby allowing a level of personalization not available before in the analysis of blood test results.

Once implemented, outlier analysis can flag for rapid changes, even within average healthy population values, and can deal with an important special case in the robustness analysis of AI methods, so-called out-of-distribution samples, that are especially critical in medical scenarios, where they can have fatal consequences. For this reason, model-driven approaches are better when they are shown to be more robust and can stand shallow adversarial attacks, such as single-pixel perturbation.

The methods for explainability implemented in our technology include perturbation and sensitivity analysis to ascertain the impact that changing the model inputs, and the models themselves, impose on their outputs. Using algorithmic probability, models are then ranked by complexity, with models whose explanatory power is in inverse proportion to their simplicity being favored. From visualizations of intermediate network layer activations and learned filters using medical literature, insights about the evolution of patient or user markers that combine all user data and medical literature in a computable form, are exploited to offer a highly informed pre-diagnosis and different disease progression and prognosis models. Traceability during the decision/

classification process can be guaranteed by the causal method through the implementation of a procedure similar to the layer-wise relevance propagation (LRP) method, but applied to non-statistical networks, to ensure that the decisions are supported by both meaningful algorithmic patterns in the input data and medical knowledge ingested by the platform. LRP operates by propagating the prediction backward in a neural network, whereas our platform propagates information back to the disease networks informing the first principles of possible progression, that is then tested against current or new data, to be adjusted and the models updated.

Explainability can also be ensured through the implementation of local interpretable model-agnostic explanations (LIME) at low levels, and at higher levels by a fully rule-based expert system that only uses deep neural networks for representation purposes. Methods and user interfaces should be designed for interpretability, explainability, and confidence, including information about possible failures, inaccuracies, and errors that should be accessible to the expert user for consideration and evaluation.

Conclusions

We propose a phased but responsible use of AI for precision medicine and value-based healthcare that is able to ensure that clinicians and patients understand the reasoning behind AI-based conclusions, enabling them to understand their chain of inference and justify their final call.

Combined, these proposals are designed to create trust and confidence in AI-based reasoning in clinicians, patients, and users, promoting responsible applications of AI that have the potential to transform medicine and healthcare.

References

1. Brainard J. Rethinking retractions, data analysis and graphics. *Science*. 2018;362(6413):390–3.
2. Tegner J, Zenil H, Kiani NA, Ball G, Gomez-Cabrero D. Perspective on bridging scales and design of models using low-dimensional manifolds and data-driven model inference. *Phil Trans R Soc A*. 2016;374(2080).

3. Zenil H, Kiani NA, Zea A, Tegnér J. Causal deconvolution by algorithmic generative models. *Nat Mach Intell.* 2019;1:58–66.
4. Zenil H, Kiani NA, Marabita F, Deng Y, Elias Y, Schmidt A, Ball G, Tegnér J. An algorithmic information calculus for causal discovery and repro-gramming systems. *iScience.* 2019;19:1160–72.
5. Zenil H, Minary P. Training-free measures based on algorithmic probability identify high nucleosome occupancy in DNA sequences. *Nucleic Acids Res.* 2019;47: e129.
6. Zenil H, Hernández-Quiroz F, Hernández-Orozco S, Saeb-Parsy K, Hernández De la Cerdá H, Riedel J. distance and colour-based scores for blood test risk stratification. 2020;bioRxiv:2020.02.09.941096. <https://doi.org/10.1101/2020.02.09.941096>.
7. Zenil H, Schmidt A, Tegnér J. Causality, information and biological computation: an algorithmic software approach to life, disease and the immune system. In: Walker SI, Davies PCW, Ellis G, editors. *Information and causality: from matter to life.* Cambridge: Cambridge University Press; 2017. p. 244–79.



AIM and the Nexus of Security and Technology

20

Kiran Heer Kaur

Contents

Introduction	294
Human Perfection and the Security Implications of Biomedical Technology	295
Eugenics: The Painful History of an Idea	295
Can “Liberal Eugenics” Really Work? The Relationship Between New Wave Eugenics and Security	296
Exploring the Symbiotic Relationship Between Technology and Security	299
Hacking Healthcare: The Violation of Patient Data	299
Concluding Thoughts	301
References	302

Abstract

Artificial Intelligence has played a key role in innovating modern medicine over the course of the twenty-first century. As a result, technology has greatly impacted and revolutionized the field of healthcare. However, despite all the progress that has been made by technology, it must be emphasized that with any form of new technology comes some degree of risk; whether this is from existing technology or from the emergence of future technologies. In order to highlight the security challenges that may derive from increased technological development, we will be drawing upon two prominent and relevant security issues. The

first issue is the development and potential use of human enhancement technologies. It will be explored how the increase of enhancement technologies may lead to further issues and inequalities within society. Drawing from historical case-studies, it will be explicitly highlighted how the use of human enhancement technologies may lead to the irresponsible misuse of technology. To further highlight how technology can increase security risks, it will also be discussed how cyberattacks can impact the healthcare industry. Looking at the use of electronic medical records (EMRs) in healthcare, it will be analyzed how existing technologies can also increase security challenges. The overall purpose of this chapter is to highlight the symbiotic relation between technology and security and to further draw attention to prominent security threats in healthcare.

K. Heer Kaur (✉)
Center for Social Development, Wolverhampton, UK

Keywords

Artificial Intelligence · Machine Learning · Natural Language Processing · Liberal Eugenics · Genetic Engineering · CRISPR/Cas9 · Big Data · Data Security and cybersecurity

Introduction

The introduction of technology, particularly Artificial Intelligence (AI) in the form of machine learning (ML), natural language processing (NLP), and deep learning (DL), have become essential and impactful tools in modern medicine. Over the course of the twenty-first century, technology has become conveniently and effectively embedded into all spheres of life through technological innovation across all sectors in society. Consequently, the innovation of technology has greatly impacted and revolutionized the field of healthcare and medicine as a whole. This is evident as currently large corporate giants like Google have created AI that can efficiently and more precisely detect lung cancer than traditional doctors, [1] machine learning (ML) AI can be used to detect schizophrenia [1], and ML and wearable technology can be used to detect the onset of heart disease [1]. The aforementioned examples are just a small selection of technological innovations that demonstrate the impact technology has had within medicine.

While many of these procedures have been beneficial for the vast majority of the population, numerous headlines and bold sweeping statements continue to make headway about the negative impact of technology and how this will be society's greatest pitfall and challenge in the future. Public influential figures such as Bill Gates and Elon Musk have expressed their concerns about the future of technology, with the latter lamenting how technology could be one of the biggest issues we combat in the coming decades. Although many of these headlines are false and plagued with misinformation and clickbait titles, it must be emphasized that with any technology and medical procedure comes

some degree of risk. Therefore, the purpose of this chapter is to provide information about the future security challenges that could derive from the use of technology in medicine, and to further explore the symbiotic relationship between security and technology in a medical context by using real life case-studies and addressing some of the most pressing global security concerns in the twenty-first century.

In order to accomplish this, two prominent and pressing security issues will be discussed at length. The first issue is human enhancement through the use of biomedical technologies. As the future of medicine continues to advance and the application of AI allows doctors, scientists, and genetic engineers to break through biological limits, using technology to enhance human beings is becoming a reality. While this topic has already been heavily discussed from an ethical, moral, and philosophical lens, the aim of this chapter is to explore the potential security and political challenges that could derive from the use of these technologies. Investigating this issue from a security and political perspective is especially important, as in the future the use of these technologies could have severe consequences upon humanity. Beyond debating whether biomedical technologies should be used to enhance human beings or not, it is important to have critical awareness and knowledge on the security challenges that could result from this technology.

The second issue that will be discussed is the use of big data in healthcare and patient privacy. As many healthcare services across the globe now use digitalized and electronic medical records (EMRs) or electronic health record (EHRs), this increasingly poses numerous risks and challenges, as it potentially leaves many healthcare institutions and practices vulnerable of cybersecurity threats. To illustrate this point further, the hacking of the Finnish psychotherapy center, Vaastamo, and the UK's National Health Service (NHS) will be analyzed in great detail. The purpose of this is to shed light on how technology can sometimes be a double-edged sword with real life implications. Although there have been many benefits from the use of technology and AI in medicine, this does not mean that all risk and

security threats are diminished. By assessing the cybersecurity issues that could derive from technology, it better equips the healthcare industry, governments, and the general public of the flaws in the system, so that these issues can be resolved.

Human Perfection and the Security Implications of Biomedical Technology

The future of medicine is one that invites positive innovation and has the potential to benefit many people across the globe. Despite this however, there are still many issues that could occur from the development and employment of such technology. This very sentiment was expressed by Francis Fukuyama, in his book *Our Posthuman future Consequences of the Biotechnology*, [2] in which Fukuyama had expressed concerns about biotechnology, and the consequences that could derive from a posthuman future. In essence what Fukuyama was trying to convey is that, as the future of medicine gets increasingly innovative and technologically adept, this could have profound consequences for humanity. This is because as more research is conducted within certain specialties, for example, within reproductive medicine and genetics, this could lead to further problems as the more our knowledge and capacity to interfere increases, the more we could alter things for the detriment of humanity. Fukuyama discusses this very phenomenon in his book, as he talks about the potential impact of biotechnology and genetic engineering, and hypothesized that it will “dominate government policy in future” [2]. Fukuyama continues by saying that in the future, “the ultimate prize of modern genetic technology will be the designer baby” [2]. A designer baby is one that has been genetically altered to have enhanced features and/or attributes.

Eugenics: The Painful History of an Idea

Throughout history, the belief that certain human beings are better than others has dominated the world. People that possessed certain skin colors, ethnicities, genders, sexualities, and were

from different social classes, that did not conform to the typical societal expectation, were victimized within society. This led to great injustices such as slavery, imperialism, colonization, racism, and sexism, throughout the twentieth century. However, it was not enough to claim someone was inferior when compared to another human being simply by the aforementioned defining factors, it needed to be proven that their genes and DNA meant that they were physically inferior.

This idea of possessing better features or attributes due to “better genes” is often referred to as eugenics [3]. The term eugenics was first coined in 1883 by Francis Galton, the cousin of Charles Darwin [3]. Galton had advocated the idea that couples should marry based on who are the fittest individuals in order to ensure a better human race, when procreating [3]. He also advocated that this idea of marrying individuals that are of a better social standing could be enforced through monetary gain provided by state governments [3]. This idea has set the groundwork for future practices of eugenics, in which selective breeding had become the “underpinnings of state sponsored discrimination, forced sterilization and genocide” [3]. The ideology surrounding Eugenics has been around since the late nineteenth century, [4] with practices occurring across the United States (US) and in Europe. In the US, scientists had worked alongside politicians to theorize and practice methods of eliminating the reproduction of those considered to be of a lesser gene pool, and thus, hereditarily weaker. Increasing reproduction and breeding among those considered to be hereditarily stronger was considered the way forward to promote a better and stronger society [4]. These early eugenic programs in the US are noteworthy to mention as they served as social models for preliminary eugenics programs in Nazi Germany [4].

The advancement of science had played a pivotal role in highlighting this phenomenon and had created further barriers and hierarchies among human beings. Using science to mark the genetic differences among people predates the twentieth century, and can be seen in Arthur De Gobineau’s, *An essay on the inequality of the human races* [5]. The significance of this is that it presents

how science, or rather pseudoscience, has been used to support and validate discriminatory agendas. This idea is crucial as it plays an integral role throughout twentieth century Europe, and led to a number of immoral experiments being conducted on human test subjects.

The most infamous and devastating examples arises from Nazi Germany, under experiments led by Josef Mengele. In Nazi Germany, people from marginalized and socially outcast groups (those of Jewish faith, Polish nationals, Roma, political prisoners, USSR prisoners, homosexuals, catholic priests, and those with physical disabilities) [6] were used as human test subjects without permission or consent, and were subjected to cruel and dangerous experimentation so that doctors could study genes.

Mengele was famous for his fascination with genetics, particularly the genes of twin siblings, leading him to conduct unethical and unorthodox practices to study that phenomenon further [7]. Holocaust survivors and fraternal twins, Irene Hizme and Rene Slotkin recount their experiences of being used as human test subjects in a 1995 interview. In the interview, Ms. Hizme recounts how she was spending a lot of time at the doctor's office and was in and out of hospital. She remembered how she was very sick after the doctors had taken blood from the left side of her neck [8]. She further discussed how she would have to wait for long periods of time to be measured and weighed, and recalled how she would often be injected, without consent, with a mysterious substance that afterwards made her unwell [8]. Likewise, Eva Mozes Kor, who was sent to Auschwitz with her twin sister Miriam, also recalls similar memories. Ms. Kors explains how "the doctors would take their blood and had given them regular injections." After one of those injections, Kors had also become extremely ill [9]. In addition to the morbid fascination and experimentation on twins, other experiments that aimed at human perfection, and forced sterilization in men and women, were also conducted. To sterilize women, the doctors had formulated an "irritating solution to be injected into the uterus" [10] against the will and knowledge of these women. These experiments are particularly important as they highlight the fundamentally

flawed concept of human perfection. To what extent is creating a "perfect human race" worth the lives of innocent people? The experiments were conducted so that the Nazi's could understand how the human body functioned, in order to use this to their advantage. By understanding how to remove and damage reproductive organs of those considered inferior, they hoped it could aid them to reach their goal of eradicating the impure and be a step closer in creating the Aryan race. The study of twin siblings was particularly sinister as doctors wanted to understand how harming one twin would have affected the other, until it led to their eventual death. By analyzing the corpses of twins, the doctors would try to look for anything that supported their hypothesis that certain people are born genetically inferior, allowing these people to then be ostracized and eradicated from society; thus, leaving behind only those considered to be pure and genetically superior. This idea, though it may seem convoluted, has great significance and relevance to the study of reproductive technologies today.

Can "Liberal Eugenics" Really Work? The Relationship Between New Wave Eugenics and Security

At present day, similar studies are being conducted to those practiced in Nazi Germany to see if there are any direct correlations between genes and certain behavioral qualities [2]. What this connotes is that studies are being undertaken to determine if there are links between genes and behavioral attributes such as "intelligence, aggression, sexual identity, criminality and alcoholism" [2]. This is of great importance as if certain genes were found to be "better" than others it begs the question of what could happen in the future if people were to take advantage of this? For example, if wealthy parents in the future could choose specific genes for their offspring and pick certain qualities and features that are deemed "better," how would this impact parents who would not be able to select genes for their offspring, due to financial restrictions? Would this create an unfair precedent in society?

Francis Fukuyama expresses his concerns that human enhancement and designer babies could become a major “political issue that may someday come to dominate politics” [2]. This is because if wealthy parents could alter the genetics of their potential offspring, they would be altering the germline not just of their offspring but also of all future descendants [2]. This could have a profound impact upon humanity, if those with better genes are to be favored or championed in society, while others are discriminated against simply for their genes. Fukuyama claims that the impact of this could someday lead to a “full scale class-war” [2].

A class war in the future, dictated by one’s genes is a haunting reminder of the ideologies and practices that emerged in the original iterations of eugenics. Despite this, some defend the use of genetic engineering and human enhancement through the use of liberal eugenics [11]. Liberal eugenics is a newer version of eugenics in which authoritarian eugenics is rejected, meaning that state governments would have no decision-making role in the process. Instead, individuals would have the liberty and freedom to decide how to utilize these technologies [11]. The fundamental principles of liberal eugenics are that individuals would be well-informed about future technologies and their utility, and thus, can make conscious decisions about how to use them [11]. Liberal eugenics is often defended by likening the choice of using biomedical and enhancement technologies, as having a series of choices in a liberal society [11]. What makes up the core foundations of a liberal society is allowing people to have the freedom and discretion to choose the life that they want and respect the choices of one another [11]. One of the core principles of liberalism itself is “acknowledging the right of others to make choices that do not appeal to us” [11]. Therefore, when applying this liberal notion to eugenics, instead of restricting the freedom of citizens through state-sponsored eugenic practices, it is returning freedom and offering individuals more choices to better their, and their future offspring’s, life. By putting the control into the hands of individuals and allowing prospective parents to make decisions in the best interest of their future offspring, it is believed that enhancement technologies would be used responsibly.

Hence the risks of creating social hierarchies and division in society would be mitigated [11].

While, the idea of liberal eugenics attempts to overcome the negative connotations attached to original eugenics practices, the theory of liberal eugenics has some fundamental flaws. The theory bases a lot of assumptions of responsible use of this technology in an idealized utopic liberal society; however, there are some important issues with this. Firstly, if biomedical technology was available and legal for global use, then it should be factored in that not every country adopts a liberal or democratic social structure. In countries that function through communism, socialism, authoritarian rule, or dictatorship systems, non-state sponsored eugenics would be difficult to facilitate. The theory of liberal eugenics is purely a Westernized concept that would only be compatible for liberal societies. Although approaching eugenics from a liberal approach could work in countries that adopt a liberal system, the theory of liberal eugenics is limiting.

This is extremely important to consider from a security context. If states that employed non-liberal systems instructed citizens on how to use these technologies or encouraged the use of genetic enhancement to create better citizens, this could lead to an array of other problems. For example, this could create chaos in the international community, if in different societies, whole populations were to genetically enhance future generations to be “smarter,” “stronger,” and “attractive,” whereas in some countries these technologies were not accessible to all citizens or were entirely outlawed. Although it is incredibly difficult to measure and predict the exact response of these technologies from different countries, it needs to be highlighted that biotechnology and the power to modify humans carries with it significant risks and challenges to the global order.

Furthermore, the concept of liberal eugenics is also flawed as it could lead to the inadvertent discrimination against certain groups of people within society. The potential of gene editing brings with it many possibilities, as it could result in the eradication of diseases and illnesses in the future. For example in 2018, media reports surfaced that doctors had illicitly used gene

editing technology to modify the genes of two twin girls [12]. The modification was intended to make the twin girls immune to human immunodeficiency virus (HIV) [12]. At the time the news broke, this caused much anger, concern, and controversy in scientific circles as the doctor responsible, He Jiankui, conducted the procedure unlawfully and unethically [12]. By using clustered regularly interspaced short palindromic repeat (CRISPR)-associated system (Cas) technology, [13] Jiankui was able to edit the gene CCR5, that is responsible for causing HIV [12]. This was achieved as the CRISPR/Cas9 system allows it to be possible to locate a certain part of DNA inside a cell that can then be edited [14]. This is possible as Cas9 acts as a pair of “molecular scissors” [15] that is able to cut DNA [15]. Within Cas9 is an enzyme, guide RNA (gRNA), that directs Cas9 to the exact part in the genome that needs to be cut [15]. Once the incision has been made, the DNA will try to repair itself; however scientists can intervene and hack this system and provide the DNA with models to replicate, thus allowing genetics engineers the ability to modify and alter genes [15]. In the case of the two twin girls, Jiankui claims he was able to effectively remove the gene that causes HIV [12].

Although this procedure was orchestrated unlawfully and widespread use of gene editing technologies are still outlawed, the success of such a procedure demands important conversations and considerations. If it is possible to eradicate diseases and genetic mutations, what would this mean for the future of disabled people? As many gene editing technologies focus on “the correction or removal off disabling genetic traits” [16], this raises some questions about whether this would be purposely eradicating those born with disabilities [16]. By trying to “fix” people that would be born with genetic mutations or anomalies, it needs to be questioned what sort of message this would send to individuals that are born and live with disabilities. While having a disability can make some aspects of life harder, it does not mean that the quality of that life cannot be as good, fulfilling, or happy as someone who was not born with a genetic mutation.

Even though liberal eugenics champions the idea of individuals picking genes for their child, this may lead to unconscious biases surfacing in the future, in which those born with disabilities face eradication in the name of modernity and progression. This itself cannot just be viewed as a moral and ethical dilemma but can also be framed as a security issue. Gene editing threatens the future for those that would be born, and currently live with, genetic mutations [16]. It reinforces negative connotations about disabled conditions and risks “reverberating [these opinions] throughout wider society” [16]. In a future where genetic editing technologies are legal in society, the security implications that could entail for different people are imperative to consider. The first wave of eugenics aimed to eradicate members in society that were deemed “less than” and genetically inferior. Liberal eugenics attempts to remedy this by promoting liberal and positive eugenics; however in doing so, this may lead to widespread unconscious discrimination.

Despite the fact that this technology is still not widely implemented within society, and that there is still a long way to go before scientists, genetic engineers, and doctors can perfect and safely use this technology, it is still important to consider the consequences at this moment in time. This is largely due to the fact that some of the technical problems that genetic engineers and doctors have encountered before can largely be resolved with AI and ML algorithms, in the twenty-first century. For example, one of the biggest logistical challenges for engineers and doctors when it came to using CRISPR/Cas9 was ensuring that if they are cutting and replacing genes, what combinations would work when replacing the original. As there can be numerous combinations, it would be difficult to accurately analyze all this information in a short amount of time. However, machine learning algorithms can largely resolve this issue as a computer can be given large datasets and analyze accurately and efficiently the best combination and outcome from previous datasets. Thus, as the reality of GE gets closer, it is imperative to consider every angle of the debate, especially with regard to security.

Exploring the Symbiotic Relationship Between Technology and Security

Over the past decades, technology has fundamentally transformed the field of medicine. Not only have pivotal developments been made through actual medical procedures such as genetic engineering as previously discussed, but technology has also changed the way patient information and data has been collected. Most hospitals and healthcare practices at present use some form of electronic medical records (EMRs) or electronic health records (EHRs) as a tool to accurately collect vast amounts of patient data [17]. EMRs are digitally kept records that contain confidential patient information recorded by healthcare professionals or personnel in a medical setting [17]. Examples of EMRs include lab results, routine check-up clinical notes, prescriptions for medication, and clinical progress reports [17]. EMRs are crucial in maintaining hospital systems and ensuring patient care [17]. The vast amount of patient information available from EMRs can be considered as a form of big data [18].

Big data are mass volumes of structured or unstructured information. In comparison to regular data, big data, as warranted by the name, is bigger in size and can be characterized by volume, variety, and velocity [19]. In recent years, big data has become increasingly important for both the private and public sector due to the important information that can be found within the huge and complex datasets [18]. The information that can be found in these datasets can then be used to study, analyze, assess, and predict trends, patterns, and even consumer behavior. In healthcare, big data is especially important as it can be used to highlight important trends in the medical field and can also be used for biomedical research [18]. This type of data is especially important as it can be used for machine learning purposes. This is because researchers in the field of machine learning (ML) have gained access “to a large quantity of high-quality medical data” [17] that can be used to teach computers about past cases to better equip the technology to recognize future cases.

The reason why this is important to note is because as the healthcare field gets increasingly more digital and employs the use of ML algorithms and other AI tools, this may lead to more complications and potential security threats in the future, since valuable data such as EMRs may be used for malicious intents and purposes.

Hacking Healthcare: The Violation of Patient Data

This very issue pertaining to the misuse of EMRs is not just a hypothetical concern as in recent years the hacking of private and confidential patient data occurred. In 2020, a scandal regarding the hacking of EMRs came to the forefront and garnered public attention after it was discovered that a hacker tried to extort psychotherapy patients after illegally obtaining access to their EMRs in Finland [20]. The Finnish psychotherapy center, Vastaamo, were at the center of the cybersecurity attack when hackers had exploited security breaches in 2018 and 2019, respectively, which had remained undetected at the time [21]. Patients had then received disturbing messages that their private records and clinical notes will be uploaded onto the internet unless they pay €500 in cryptocurrency [21]. The hackers had sent more than tens of thousands of emails to patients demanding money in exchange for their silence [20]. Sensitive information, including confidential clinical notes and social security numbers, were threatened to be published.

Indeed, a similar cybersecurity attack also affected the British healthcare service in 2017, after the UK’s National Health Service (NHS) fell victim to the global WannaCry ransomware attack [22]. The WannaCry attack brought the NHS to a halt after affecting hospitals, practices, and GP surgeries throughout England and Scotland; with thousands of appointments and operations being cancelled [22]. Similarly to the Finnish case, hackers demanded monetary gains in the form of cryptocurrency in exchange for silence over confidential EMRs that they acquired. The hackers had managed to achieve

this large-scale attack by exploiting a preexisting vulnerability in the software. Although the NHS was not a particular target in the mass global attack, the software had left them susceptible to the attack [22]. The NHS was particularly vulnerable to the attack as the NHS at the time used archaic Microsoft systems IT systems thus increasing the likelihood of cyberattacks affecting the NHS altogether [22].

The reason why these cases are important to reference is two-fold, and shall be discussed. Firstly, as patient data continually becomes more digitalized through the use of EMRs/EHRs and is used for biomedical research, ML, or any other use, it must be imperative that patient privacy and data security is ensured and maintained at all times. This is of vital importance as a violation of such personal information could result in a loss of confidence from patients, which could have a profound impact on both the patient and healthcare system. The second reason that this should be considered is because any new technology comes with some form of risk. The two cases in Finland and the UK highlight the dangers of cyberattacks and hackers in the healthcare industry. It explicitly emphasizes the vulnerabilities of technology and the long-lasting impact that it can have. While technology can help to innovate healthcare, it is important to consider that with technology comes certain risks, thus these risks should be examined in a bid to mitigate any potential catastrophes.

In the Finnish and UK case studies mentioned, the violation of patient data presents an emerging human security issue. This is because the theft, extortion, and blackmail as depicted in the Finnish case can have a profoundly negative impact on individuals that threatens their sense of security and trust. In a world where data is considered the most precious commodity becoming even more lucrative than oil [23], the private data of patients and EMRs are incredibly invaluable. The medical data of an individual can reveal private information about a patient that they may not want to disclose publicly, such as illnesses, health conditions, and mental health concerns. The theft of medical data is especially disconcerting as EMRs and EHRs have a longer life compared to other forms of personal data [24]. Compared to other forms of personal data like an individual's

online digital footprint or financial data including credit card and bank details, the lifespan of medical data is imaginably longer as the medical history of a person does not ever change [24], in comparison to other forms of data that may be subject to change. As a result, evidence indicates that stolen EMRs and EHRs are "often sold and resold on the dark web," [24] perpetrating an endless cycle. This makes the theft of EMRs especially unique and dangerous as it has the potential to cause the most lasting damage for individuals. This can lead to further attempts of blackmail, extortion, and fraudulent behavior.

Moreover, the long-term impact of stolen patient information can substantially affect and worsen a patient's health and well-being. The distress of the violation may cause trauma to patients who are contacted by hackers to pay a lump sum of money in exchange for their information. In the Finnish case, patients were reported to have "flooded victim support services" [25] due to the distress of the events. In addition to the trauma faced by patients, this may also encourage a lack of trust in healthcare services and a distrust for technology. Not only would this fundamentally affect patients but it would also impact healthcare services if people do not feel safe and entrust healthcare workers with their medical concerns. This may lead to people not attending routine check-ups and appointments that could be important in ensuring that people stay healthy, and detect the onset of any illnesses or underlying health conditions earlier. As EMRs are increasingly being used to train ML algorithms to detect and identify previous case studies to aid new cases, it is especially essential to ensure the data privacy of patients. Before implementing new technological processes in healthcare, the existing cracks from using technology first need to be resolved.

While the use of technology in medicine has shown increasing benefits it cannot be denied that this also brings a whole lot of challenges and security concerns. As mentioned, the theft of patient data is arguably one of the biggest consequences that could derive from the lack of cybersecurity provisions in healthcare. The relationship between security and technology itself is an interesting albeit turbulent one. As the world becomes more digital and adopts a digital infrastructure,

this increases the likelihood of potential cyber-attacks. Likewise, as the healthcare industry continues to utilize technology, this may also bring with it some unwarranted risks. It is therefore important to explore the link between cybersecurity and healthcare and resolve the vulnerabilities in order to make systems secure.

The healthcare industry appears to be at a higher risk from a cyberattack than any other industry or sector [26]. The impact of a cyberattack not only risks the sensitive data of patients, but it also risks infecting the infrastructure of healthcare practices, costs the health service millions, disrupts workers from doing their job, and affects patient care. Healthcare is suggested to be an attractive target for two reasons. The first reason is that the healthcare industry has an unlimited supply of invaluable data, and the second reason is that protections and defenses against cyberattacks are weak [27]. This makes the target of healthcare that much easier to exploit as a fast way to make monetary gain. Despite the impact that a cyberattack may have on the healthcare industry, evidence suggests that the healthcare industry is still unprepared with the recent 2020 case in Finland. As a result, not only are technological developments needed to improve the field of healthcare but so are enhanced security measures.

Stronger measures such as blockchain systems may be beneficial to employ, as well as other preventative measures. Blockchain is a system that is used to securely store and share information [28]. Blockchain works as data is stored in decentralized transparent blocks that are linked together by cryptographic hashes [28]. Blockchain has the potential to remedy many of the cybersecurity issues the healthcare industry faces, as it is a more secure way to collecting and storing confidential data. Blockchain, with a combination of other security measures such as encrypted emails, adequate staff training to recognize fraudulent and phishing emails, and impenetrable networks, offer increased security.

By identifying the susceptibilities that may derive from increased technological development, this will crucially benefit the healthcare industry. The relationship between technology and security is a symbiotic one as the two unequivocally go

hand in hand. A lack of security allows for technology to be exploited. As medicine continues to advance, the security provisions also need to be held to the same standard and accountability.

Concluding Thoughts

It is better to have principles covering situations that turn out to be impossible than to have no principles for situations in which we suddenly find ourselves [11]

- Nicholas Agar, Liberal eugenics: In defense of human enhancement [11].

The continued implementation of technology in medicine will certainly elevate healthcare and modern medicine. Although, a lot of these breakthroughs will have a positive impact on humanity and within medicine, it is still imperative to explore the relationship between technology and security. This is because sometimes with technology there is still a degree of risk that could derive from the use of it. This was explicitly highlighted throughout this chapter as despite the advances that have been made, there are still some fundamental security concerns.

As AI continues to advance medical practices and paves the way for greater innovation, it is important to critically analyze the purpose of technology and ensure its responsible and safe use. This entails analyzing potential security issues before they arise. In terms of GE, this technology has the power to greatly benefit a lot of people; however, equally it can also bring about a lot of harm if used irresponsibly. Thus, it is wise to have as Agar suggested, security measures for situations that could happen as opposed to no measures or considerations at all. Genetic enhancement has the potential to be misused; hence it would be wise to critically analyze how this could occur and mitigate the possibility of misuse before it is a widely normalized part of society.

Likewise, when analyzing technology and the negative impact that it could have, it is also wise to investigate existing uses of technology and ensure that they are being used safely. As the healthcare industry continually gets more digitalized, it is also important to cover all bases, and ensure

there are adequate security provisions in place. By closely examining the relationship between technology and security, we optimize our chances of having a safe digital future.

References

1. Jack C. Top 6 AI breakthroughs in healthcare. <https://www.docwirenews.com/docwire-pick/the-top-6-ai-breakthroughs-in-healthcare/>. Accessed 11 Nov 2020.
2. Fukuyama F. Our posthuman future: consequences of the biotechnology revolution. London: Profile Books; 2003.
3. Farber SA. US Scientists' role in the Eugenics Movement (1907–1939): a contemporary biologists perspective. *Zebrafish*. 2008;5(4):243–5.
4. Grodin MA, et al. The Nazi physicians as leaders in eugenics and “Euthanasia”: lessons for today. *AJPH*. 2018;108(1):53–7.
5. de Gobineau A. The inequality of human races (trans: Collins A). 1915; London: Heinemann. p. 205–12.
6. United States Holocaust Memorial Museum. Medical experiments. 2020. <https://www.ushmm.org/collections/bibliography/medical-experiments>. Accessed 16 Nov 2020.
7. Meyers W. The abuse of man: an illustrated history of dubious medical experimentation. New York: Ardor Scribendi Ltd; 2003. p. 268.
8. Interview with Mr. Rene Slotkin and Ms. Irene Hizme, 19th April 1995.
9. Kors E. This questionnaire with my personal answers (Candles Holocaust Museum and Education center). https://candlesholocaustmuseum.org/file_download/inline/2ffd2717-e5e9-4b64-9e92-5361aa0fcbe1. Accessed 27 Nov 2020.
10. Blacker CP. Eugenic experiments conducted by the Nazis on human subjects. *Eugen Rev*. 1952;44:9.
11. Agar N. Liberal eugenics: in defense of human enhancement. Oxford: Blackwell Publishing; 2008.
12. Raposo VL. The first Chinese edited babies: a leap of faith in science. *JBRA Assist Reprod*. 2019;23(3):197–9. <https://doi.org/10.5935/1518-0557.20190042>. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6724388/>. Accessed 29 Nov 2020.
13. Liang P, et al. CRISPR/Cas9-mediated gene editing in human triploid zygotes. *Protein Cell*. 2015;6(5):363–72. <https://doi.org/10.1007/s13238-015-0153-5>. <https://pubmed.ncbi.nlm.nih.gov/25894090/>. Accessed 28 Nov 2020.
14. New Scientist. What is CRISPR? <https://www.newscientist.com/term/what-is-crispr/>. Accessed 28 Nov 2020.
15. NOVA PBS Official. September 9th 2020. The realities of Gene editing With CRSIPR | NOVA | PBS. YouTube Video. https://www.youtube.com/watch?v=E8vi_PdGrKg. Accessed 16 Dec 2020.
16. Boardman F. Human genome editing and the identity politics of genetics disability. *J Community Genet*. 2019;11:125–7.
17. Shinozaki A. ‘Electronic medical records and machine learning in approaches to drug development. <https://www.intechopen.com/books/artificial-intelligence-in-oncology-drugdiscovery-and-development/electronic-medical-records-and-machine-learning-in-approaches-to-drugdevelopment>. Accessed 24 Dec 2020.
18. Dash S, et al. Big Data in healthcare: management, analysis and future prospects. *J Big Data*. 2019;6(54):2–25.
19. OECD. The impact of Big Data and Artificial Intelligence (AI) in the insurance sector. <https://www.oecd.org/finance/The-Impact-Big-Data-AI-Insurance-Sector.pdf>. Accessed 28 Dec 2020.
20. Politico. Hackers seeks to extort Finnish Mental Health patients after data breach. <https://www.politico.eu/article/cybercriminal-extorts-finnish-therapy-patients-in-shocking-attack-ransomwareblackmail-vastaamo/>. Accessed 28 Dec 2020.
21. Gromek M. Ransom Hackers in Finland are using psychotherapy medical record as ammunition. <https://www.forbes.com/sites/michalgromek/2020/10/31/ransom-hackers-in-finland-are-usingpsychotherapy-medical-records-as-ammunition/?sh=98b22df2f2eb>. Accessed 28 Dec 2020.
22. Acronis International GmbH. The NHS Cyber Attack. <https://www.acronis.com/en-gb/articles/nhs-cyber-attack/>. Accessed 29 Dec.
23. The Economist. The world's most valuable resource is no longer oil, but data. <https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data>. Accessed 30 Dec 2020.
24. Conaty-Buck S. Cybersecurity and healthcare records: tips for ensuring patient safety and privacy. *Am Nurse Today*. 2017;12(9):62–5.
25. The Guardian. Shocking hack of psychotherapy records in Finland affect thousands. https://www.theguardian.com/world/2020/oct/26/tens-of-thousands-psychotherapy-records-hacked-infinland?CMP=Share_AndroidApp_Other. Accessed 2 Jan 2020.
26. Martin G, et al. Cybersecurity and healthcare: how safe are we? *BMJ*. 2017;358:j3179.
27. Coventry L, Branley D. Cybersecurity in healthcare: a narrative review of trends, threats and ways forward. *Maturitas*. 2018;113:48–52.
28. Chen HS, et al. Blockchain in healthcare: a patient centered- model. *Biomed J Sci Tec Res*. 2019;20(3):15017–22.

Bibliography

United States Holocaust Memorial Museum, Oral Interview with Mr. Rene Slotkin and Ms. Irene Hizme. <https://collections.ushmm.org/search/catalog/irn504815>. Accessed 26 Dec 2020.



AIM in Unsupervised Data Mining

21

Luis I. Lopera González, Adrian Derungs, and Oliver Amft

Contents

Introduction	304
Association Rule Mining	305
FRM	305
BRM	306
Likelihood Mining Criterion (LMC)	307
LMC–FRM Comparison	307
Basic Rule Mining Example	308
Methodology	308
Evaluation	309
Census and Chemical Exposure Database Mining	310
Methodology	310
Evaluation	311
Rehabilitation Routine Mining	312
Methodology	312
Evaluation	312
Conclusions	314
References	315

Abstract

This chapter explores the differences between association rules extracted using the likelihood mining criterion (LMC) and rules extracted by using frequent item-set rule mining (FRM). LMC provides a change in perspective for rule selection, from a measure of frequency in the dataset to a measure of relationship between the rule items. For illustration, this chapter presents the evaluation of qualitative differences between LMC and FRM rules with three examples: (1) a basic rule mining

L. I. Lopera González (✉) · A. Derungs
Friedrich Alexander University Erlangen-Nuremberg,
Erlangen, Germany
e-mail: luis.i.lopera@fau.de; adrian.derungs@fau.de

O. Amft
Digital Health, Friedrich-Alexander University Erlangen-Nuremberg, Erlangen, Germany
e-mail: oliver.amft@fau.de

scenario to illustrate LMC properties, (2) an analysis relating socioeconomic information and chemical exposure data, and (3) mining behavior routines in patients undergoing neurological rehabilitation. Results show that LMC is capable of extracting rare rules and does not suffer from support dilution. Furthermore, LMC focuses on the individual event generating processes, while FRM focuses on their commonalities.

Keywords

Likelihood · Likelihood mining criterion · Min-support · Association rule mining · Frequency rule mining · Bayesian rule mining · Support dilution · Mining medical datasets · Routine mining

Introduction

Association rules can model a process by describing the relationship between variables. In a dynamic process, for example, a rule states that a change on an input will cause a change on an output. As the process evolution is stored in a dataset, association rule mining (ARM) can extract the original relationship between the process inputs and outputs. The common approach for ARM is frequent association rule mining (FRM). Rules extracted by FRM, e.g., $X \rightarrow Y$, have support and confidence greater than a user-specified min-support and min-confidence thresholds [1]. Formally, support of an item-set X is defined as the number of transactions T in the database D containing item-set X divided by the number of transactions in the database. Equation 1 shows the support of a rule $X \rightarrow Y$. The rule confidence is defined using support in Eq. 2.

$$Supp(X \rightarrow Y) = \frac{|\{(X \cup Y) \subseteq T_k, (X, Y) \neq \emptyset, \forall T \in D\}|}{|D|} \quad (1)$$

$$Conf(X \rightarrow Y) = \frac{Supp(X \rightarrow Y)}{Supp(X)} \quad (2)$$

The following thought experiment illustrates one of FRM's limitations. Suppose all supermarket receipts from the winter holidays are used for

rule mining. One would expect to see rules that associate the ingredients used for winter holiday meals. Now consider a year worth of receipts from the same supermarket. In the light of the additional data, the winter holiday meal ingredients would not have enough support to be extracted. In other words, the minimum support (min-support) threshold used to extract rules in a small dataset will not work in an extended version of the dataset due to the definition of rule support. In this chapter, min-support's dependency on the dataset size is called support dilution.

A dataset can be viewed as the collection of items sampled from multiple generating processes. However, FRM has the implicit assumption that all items are generated at the same rate. In practice, processes can generate items at different rates. For example, people in Germany occasionally buy white sausages, but when they do, they buy wheat beer too. So the rule “if white sausage then wheat beer” is a rare rule when compared to frequent rules, e.g., “if milk then eggs.” FRM can extract rare rules by using a low min-support threshold. Unfortunately, spurious item associations may create unwanted rules that FRM's threshold cannot eliminate. Filtering out unwanted rules for FRM has been addressed in the past [2, 3]. However, the processing required to separate rules does not generalize [4].

The field of medicine advances by capturing evidence that supports a given hypothesis. As the means to collect data surpasses the current capacity to analyze it, automation is required to extract new knowledge. ARM as a tool provides investigators the ability to sift through data repositories automatically. However, FRM methods are inadequate to extract useful knowledge from medical repositories, if the interest is association quality and not how frequent a certain observation occurs in a dataset. Therefore, this chapter reviews the use of likelihood as a means to select rules with quality measured independently of the dataset size. In particular, the minimum likelihood mining criteria (LMC) is explored and compared to min-support FRM. LMC has been proposed to replace min-support as primary rule selection criteria in ARM [5].

This chapter is organized as follows: section “[Association Rule Mining](#)” presents an overview of ARM algorithms. Section “[Likelihood Mining Criterion \(LMC\)](#)” describes LMC and illustrates

rule selection differences between FRM and LMC by extracting all atomic rules, i.e., rules of the form $X \rightarrow Y$, where $|X| = |Y| = 1$, from five datasets commonly used in ARM literature. Sections “[Basic Rule Mining Example](#)” through “[Rehabilitation Routine Mining](#)” illustrate LMC and FRM in three ARM applications: a synthetic timeseries, a dataset linking socioeconomic variables with health-related chemical exposure information, and a dataset of daily behavior routine annotations of patients with hemiparesis. In each setting LMC and FRM extracted rules are compared and evaluated for quality and usefulness.

Association Rule Mining

ARM algorithms can be generalized to have two stages: (1) search for useful item-sets, and (2) search for adequate rules within these item-sets [6–8]. In the context of ARM, useful item-sets are categorized as passing a min-support threshold and adequate rules depended on a secondary rule selection criterion. As the problem of support dilution became apparent [9], many approaches looked to circumvent the min-support threshold or to define new rule interest metrics, but always relied on support properties to search the lattice of rule candidates. In this section, algorithms that rely on support to select rules are grouped under FRM methods.

Conceptually, there is no correlation between rule frequency and rule interest, as pointed out by Li et al. [10]. Therefore, alternative rule interest metrics have been proposed, inspired by probability

theory, where rules are created based on the relationship between the rule and its conforming items. In this section, algorithms that use probability theory to select rules are grouped under Bayesian rule mining (BRM) due to their use of Bayesian concepts to select interesting rules. LMC is considered a primary rule selection criteria class under BRM. Figure 1 illustrates a summarized taxonomy of ARM algorithms, by mining methodology, primary rule selection criteria, and search type.

The rest of this section provides a brief overview of ARM literature grouped into FRM- and BRM-based approaches.

FRM

High-utility item-set mining is the general approach to extracting useful item-sets. The idea behind high-utility item-set mining is that a utility is given to each item in a transaction. Then, high-utility item-sets are extracted by summing the utility of each item-set across all transactions and comparing them to a min-utility threshold [11]. The advantage of high-utility item-set mining is that items bought occasionally, e.g., white sausages and wheat bear, usually have high utility and are extracted. It has been shown that frequent item-set mining is a special case of high-utility item-set mining, where the transaction utility per item is one and min-utility threshold becomes min-support [12]. Nguyen et al. [13] illustrate multiple approaches to association rule mining using high-utility item-set mining. The challenge with high-

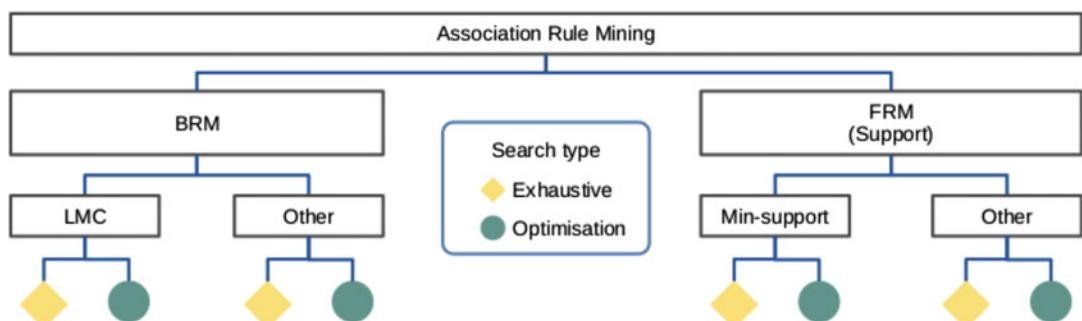


Fig. 1 Extended taxonomy of association rule mining. FRM and BRM rule mining methods represent different concepts related to the rule selection process. FRM employs support properties while the yet less established

BRM uses probabilities. In this chapter, LMC and min-support FRM are compared. LMC is considered a primary rule selection criteria class under BRM

utility mining is that the utility of an item is not always available as in the supermarket case. Therefore, the chapter's scope is limited to compare ARM based on LMC with FRM without utility.

In general, the two-stage approach to ARM requires min-support as primary rule selection criterion, and a secondary selection criterion such as min-confidence threshold. Seminal work in ARM, like the Apriori [1] or ECLAT [14] algorithms, introduced the min-confidence threshold as the main mechanism for creating association rules. As ARM was used to solve real-world problems, the threshold on confidence was producing rules with no prediction power [15]. Therefore, several interest metrics were introduced, such as lift [16], conviction [17], relative confidence [18], information gain [19], and others [20]. Alternatively, rule grammars and machine learning approaches have been presented as alternatives of the downward-closure property of support to restrict the rule search space, and extract more interesting rules. For example, Padillo et al. [21] used rule grammar and map reduce to optimize the mining process in large datasets. In timeseries, Guillam-Bert et al. [22] proposed the TITARI algorithm. They used decision trees to improve rule description and temporal specificity. However, these interest metrics were used as secondary rule selection criterion, and depended on item-sets extracted using a min-support threshold.

Rare rules are defined as the rules with support between a min-rare-support and min-support thresholds [23], where $\text{min-rare-support} < \text{min-support}$. In the rare-rule community, the available work focuses on resource-efficient extraction. For example, Tsang et al. [24] proposed the RP-Tree structure to facilitate rare rule extraction. Liu and Pan [3] proposed RP-VRIM, an extension RP-Tree that uses vertical layout to reduce dataset scans. Borah and Nath [9] proposed the SSP-Tree that enables the search of frequent and rare pattern combinations simultaneously and considers dynamic datasets.

Selecting the correct value for min-support is difficult. Thus, several approaches proposed rule search mechanisms that omit the min-support threshold. For example, *OPUS* [25] is a frequent item-set exhaustive search algorithms that does not use the downward-closure property to traverse

the item lattice. Webb [26] presented an extension that converted *OPUS* into an ARM algorithm. Bashir et al. [6] proposed an exhaustive search algorithm for frequent item-set mining, which starts by selecting n item-sets. Then, their algorithm selects the smallest support values of the chosen item-sets, and prunes the search space using the downward-closure property. Fournier-Viget and Tseng [27] proposed the TNR algorithm where the top n nonredundant rules were extracted. Both TNR and Bashir et al.'s algorithms required min-confidence to be specified.

Support dilution is a common ARM problem in dynamic databases. Cheung et al. proposed FUP [28], the first algorithm to consider efficiently updating extracted knowledge after adding new transactions to a database. Tobji and Gouider [29] extended FUP for different user-given support thresholds after adding transactions. Aqra et al. [30] proposed the Aprior algorithm to improve rule maintenance under append, update, and remove operations over the dataset. All these methods have in common that the min-support threshold needs to be manually updated in order to maintain interesting rules.

Although Tian et al. [31] proposed the use of probabilistic metrics of rule interest based on Bayes theorem, their methodology is based on a frequent item-set extraction, and proposes the Bayesian confidence and Bayesian lift as secondary rule selection criteria.

BRM

Bayesian methods for ARM are still less frequently investigated, nevertheless offer elegant features and complementary properties to FRM. The following review summarizes key contributions to BRM.

Li et al. [10] introduced local support as primary rule selection criteria in medical datasets. Local support is an approximation of the likelihood probability based on data observations. Gay and Boullé [32] have used Bayes and Information theory to select rules with the best classification power. Their proposed *level* sets a boundary for extracting interesting rules, with a preference for simpler models, i.e., shorter rules. However, the

level does not have a monotonic or anti-monotonic property that can be exploited to minimize the search space. Therefore, their methodology looks for locally optimal rules. In contrast to *level*, LMC has the anti-monotonic property [5], which reduces rule search space. Lopera G. et al. [33, 34] used increasing belief in the recursive application of Bayes theorem as rule selection criterion. Increasing belief can be simplified to a threshold on the likelihood. LMC differs from increasing belief by using a fixed minimum likelihood rather than the rule's premise probability as threshold value.

When LMC is used to replace min-support in FRM algorithms, existing interest metrics can be used as secondary rule extraction criteria, including Bayesian confidence and lift. As LMC extracts rules from the entire support range, min-support and min-rare-support thresholds can be defined after mining rules to label LMC rules as rare, e.g., following the FRM definition of rare rules. Although LMC requires data representations like SSP-Tree to maintain item-set counts, in dynamic datasets, LMC does not require the maintenance of thresholds between dataset updates. Unlike min-support, LMC does not depend on the dataset size.

Likelihood Mining Criterion (LMC)

When considering data supporting a hypothesis, Bayes' theorem is a useful tool to measure the change in belief when new data becomes available. In general, Bayes' theorem states the relationship between posterior and likelihood scaled by the prior of the data and the hypothesis. The posterior is defined as the probability of the hypothesis H being true given the data D . The likelihood is the probability of observing D given that H is true. The priors are the respective probabilities of D and H . Equation 3 illustrates the relationship.

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} \quad (3)$$

The ARM task can be restated as finding the association such that an item-set D provides evidence to an item-set H in the form $D \rightarrow H$. Using

this new formulation, and approximating probabilities from data, it follows that confidence is equivalent to the posterior $P(H|D)$, and local support is equivalent to the likelihood $P(D|H)$. In this chapter, LMC is designed to find atomic rules when the likelihood is at least 50%, describing associations that are better than random choice for the item in H . Equation 4 illustrates LMC for a rule of the form $a \rightarrow b$, where a and b are items in the dataset.

$$\text{LMC} : P(a|b) \geq 0.5. \quad (4)$$

LMC–FRM Comparison

Difference between FRM and LMC mined rules were evaluated using five datasets commonly used in ARM literature: the Chess, Accidents, Retail, Mushrooms, and T40I10D100K available at <http://fimi.ua.ac.be/data/>. A miner extracted all possible atomic rules from each dataset, and illustrates which rules pass FRM and LMC criteria. To extract all atomic rules, the miner scanned each dataset once and recorded all pairwise combinations, keeping counts of rule appearances, and individual item appearances. For example, given a transaction $T_k = \{a, b\}$, atomic rules $a \rightarrow b$ and $b \rightarrow a$ were extracted. Subsequently, rules were selected by applying LMC after each dataset was scanned. Any possible rule in the dataset follows that $\text{Conf}(r) \geq \text{Supp}(r)$, as the $\text{Supp}(X) \leq |D|$ generating a triangular rule region in the support/confidence plane. Figure 2 illustrates the triangular rule region with the boundary support = confidence. Furthermore, Fig. 2 illustrates the FRM region delimited by the min-support, min-confidence thresholds, using the following values for best visualization: min-confidence = 0.5, min-support = 0.1, and min-support = 0.002 for the Retail and T40I10D100K datasets. Fig. 2 also shows all the atomic rules available in each dataset, and highlights rules that pass LMC. Depending on the dataset, LMC can extract rules from any part of the triangular rule region, where FRM is bound to the rules made available by the min-support and min-confidence thresholds. Although all LMC rules might not be of use, the

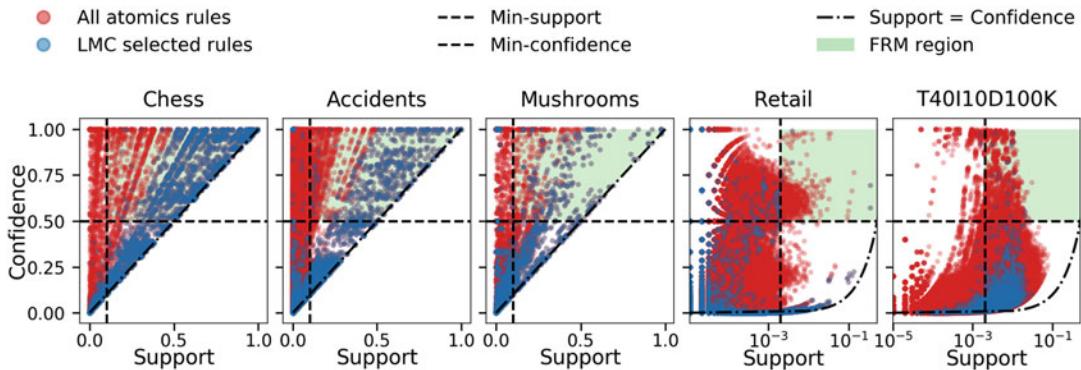


Fig. 2 Distribution of all possible atomic rules in each dataset. Rules in the FRM region pass the min-support and min-confidence thresholds. LMC rules are highlighted, and depending on the dataset, LMC rules can come from any

location on the triangular rule region. Log scale is used to display the support axis for the Retail and T40I10D100K datasets

exploration of space provides the user with information about the application and helps to guide the knowledge extraction process, and recorded all pairwise combinations, keeping counts of rule appearances, and individual item appearances. For example, given a transaction $T_k = \{a, b\}$, atomic rules $a \rightarrow b$ and $b \rightarrow a$ were extracted. Subsequently, rules were selected by applying LMC after each dataset was scanned. Any possible rule in the dataset follows that $\text{Conf}(r) \geq \text{Supp}(r)$, as the $\text{Supp}(X) \leq |D|$ generating a triangular rule region in the support/confidence plane. Figure 2 illustrates the triangular rule region with the boundary support = confidence. Furthermore, Fig. 2 illustrates the FRM region delimited by the min-support, min-confidence thresholds, using the following values for best visualization: min-confidence = 0.5, min-support = 0.1, and min-support = 0.002 for the Retail and T40I10D100K datasets. Figure 2 also shows all the atomic rules available in each dataset, and highlights rules that pass LMC. Depending on the dataset, LMC can extract rules from any part of the triangular rule region, where FRM is bound to the rules made available by the min-support and min-confidence thresholds. Although all LMC rules might not be of use, the exploration of space provides the user with information about the application and helps to guide the knowledge extraction process.

Basic Rule Mining Example

This example compares LMC and FRM miners. The goal is to cluster items from a synthetic timeseries according to their generating process, including the extraction of rare rules.

Methodology

A timeseries generator simulates processes emitting common and rare items. Using an observation window of fixed size, two miners extracted all atomic rules created by pairing the first item of the observation window with all remaining items. Then, each miner is applied a primary and secondary rule selection criteria and graphs are created using the resulting rules. When independent subgraphs formed, each subgraph was considered an item cluster that represented a generating process. For primary rule selection criteria one miner used LMC. For comparison, the FRM miner used a min-support threshold that was set using heuristics to 0.1. The selected heuristic assumed all items were equally likely to appear, i.e., 1/7, and rounding down to one decimal point.

The timeseries generator mixed two processes: (1) a common process $p_r(t)$ that frequently emitted random items and (2) a rare process $p_c(t)$ that occasionally emitted a specific pattern of items

denoted as a chain. The process $p_r(t)$ sampled vocabulary $V_r = 0, 1, 2, 3$ using a uniform distribution. In contrast, $p_c(t)$ used vocabulary $V_c = 10, 11, 12$ to emit the chain $10 \rightarrow 11 \rightarrow 12$. Chain items were always emitted in the same order, but the timing between items varied uniformly, sampled from the integer interval [1, 10] items. The timeseries generator filled the gaps between $p_c(t)$ emissions with items from $p_r(t)$, resulting in a dense timeseries. Additionally, the timeseries generator used 1000 sampled items from $p_r(t)$ and 20 chains from $p_c(t)$. $p_c(t)$ chains were uniformly distributed throughout the timeseries and could not overlap. Equation 5 shows an excerpt of a generated timeseries ts , with the items emitted from $p_c(t)$ highlighted.

$$ts = [\dots, 0, 2, 3, \mathbf{10}, 2, 0, 0, \mathbf{11}, 2, 2, 1, 1, \mathbf{12}, 2, 2, 2, \dots] \quad (5)$$

Following FRM's two-step process for rule mining, to improve the subgraph separation, the FRM miner used the following secondary rule selection criteria: (1) a min-confidence threshold of 0.5, matching the threshold on $P(a|b)$ defined in LMC, (2) a selection criterion based on the Bayesian factor, and (3) selecting rules with the highest confidence for each conclusion. Except for the min-confidence threshold, the secondary rule selection criteria were chosen to avoid additional thresholds. Equation 2 shows the estimation of rule confidence where $X = \{a\}$ and $Y = \{b\}$ confidence. The Bayesian factor was estimated using Eq. 6, where the rule $a \rightarrow b'$ denotes atomic rules in the miners final rule set A that have premise a and items other than b as conclusion.

$$\begin{aligned} \frac{P(b|a)}{P(b'|a)} &= \frac{\text{Supp}(r)}{\text{Supp}(a \rightarrow b')} \\ \text{Supp}(a \rightarrow b') &= \sum_{\forall a \rightarrow x \in A, x \neq b} \text{Supp}(a \rightarrow x) \end{aligned} \quad (6)$$

Evaluation

In accordance to the generator's characteristics, performance was evaluated by grouping extracted

rules into the following categories: (1) R_r contained all possible atomic rules that use V_r items, ($|R_r|: |V_r|^2 = 16$) (2) R_c contained all atomic, time-ordered, decompositions of the chain $10 \rightarrow 11 \rightarrow 12$, i.e., $10 \rightarrow 11$, and $11 \rightarrow 12$, ($|R_c|: 2$) (3) R_{rc} contained atomic rules of the form $i \rightarrow j$, where $i \in V_r$ and $j \in V_c$ ($|R_{rc}|: |V_r| * |V_c| = 12$) (4) R_{cr} contained atomic rules of the form $j \rightarrow i$, where $i \in V_r$ and $j \in V_c$ ($|R_{cr}|: |V_r| * |V_c| = 12$), and (5) R_{cv} contained atomic rules which were created from all possible pairwise combination of V_c items and are not in R_c , ($|R_{cv}|: |V_c|^2 - 2 = 7$).

Process separation occurred when miners only extracted rules from the categories R_r , R_c , and R_{cv} , as no rules bind items from the two generating processes. Rules in R_{cv} were not generated by $p_c(t)$. Thus, they are considered a separate category. Rules in R_{cr} and R_{rc} connect the items from $p_c(t)$ and $p_r(t)$ and no process separation is possible. R_{cr} and R_{rc} were defined as independent categories to evaluate LMC's effect on item association between frequent and rare items, when considering their position in the rule. The mining performance metric quantified the extraction rate for a rule category R with size $|R|$ as shown in Eq. 7, where A is the mined rule set.

$$\text{Extraction rate} = \frac{|\forall r \in A \cap R|}{|R|} * 100[\%] \quad (7)$$

A parametric search looked for observation window sizes in the range between [2, 500] items in one item increments. One hundred timeseries were generated for each observation window size. The following evaluation steps used a window size that minimized the chances of extracting rules from categories R_{rc} , R_{cr} , and R_{cv} . The selected window size also ensured that the miners always mined all rules from the R_c and R_r categories. The secondary rule selection methods required the generation of a new batch of one hundred timeseries. Each miner processed the new timeseries and produce results using the secondary rule selection criteria. The performance evaluation measured the average number of times items were correctly separated into generating processes $p_r(t)$ and $p_c(t)$, respectively.

The evaluation found that the FRM miner could not retrieve the items from V_c as their support was around 0.01. Therefore, any atomic rule from $p_c(t)$ will also not pass the min-support threshold of 0.1 due to support's downward-closure property. In addition, sampling extra items from $p_r(t)$ caused $p_c(t)$ chain's support to dilute. In contrast, LMC does not depend on the number of items in the timeseries. Therefore, LMC always retrieved the $p_c(t)$ chain.

Figure 3 illustrates how the rule categories were extracted as a function of the observation window size. Using an observation window larger than the expected item timing of $p_c(t)$ of five samples, LMC always extracted all rules from the R_r and R_c categories, which are needed to separate items according to their generating processes. Rules in R_{cv} were extracted when at least two partial chains were seen by the observation window. An observation window size in the range [5, 8] prevents the extraction of R_{cv} rules. Rules from R_{rc} meet LMC because the rule support is similar to the conclusion support. Thus, $P(X|Y) \approx 1 > 0.5$ and R_{rc} rules are always extracted. In contrast, R_{cr} will never pass LMC and are ignored because the rule support is smaller than the conclusion support, $\approx 20/250$, which illustrates that under LMC, rare premises cannot associate with frequent conclusions. In the LMC miner, the generating process separation task needed to remove

R_{rc} rules from the final set A using a secondary rule selection criterion. The min-confidence threshold always correctly separated the items into generating processes. The Bayesian factor only selected $p_c(t)$ rules and no items from $p_r(t)$ are grouped. Finally, the best confidence per conclusion criterion failed to separate items into two generating processes, because $p_c(t)$ item 10 is always associated as conclusion with $ap_r(t)$ item.

Census and Chemical Exposure Database Mining

The following example compares LMC and FRM miners in a database mining task. The database is the publicly available dataset from Huang et al. [35]. The dataset comprises US census tract information from the American Community Survey (ACS) 5-year summary files for the 2010–2014 period. Moreover, the dataset contains health-related chemical exposure data generated from the 2011 National-Scale Air Toxics Assessment (NATA), specifically air pollutant exposure concentration.

Methodology

Huang et al. reported results for two mining scenarios: (1) mining rules using the socioeconomic variables as premises and chemical exposure variables as conclusions ($S \rightarrow C$), and (2) mining rules within the socioeconomic dataset ($S \rightarrow S$). Huang et al. used min-support of 0.1 and lift ≥ 1 as thresholds for FRM-based rule selection. Here their atomic rule findings are replicated with an exhaustive search, min-support, and lift FRM algorithm and compared the extracted rules to LMC results. Lift is defined in Eq. 8

$$\text{Lift}(X \rightarrow Y) = \frac{\text{Supp}(X \rightarrow Y)}{\text{Supp}(X) \times \text{Supp}(Y)} \quad (8)$$

Huang et al. categorized variables as follows: socioeconomic scores were divided into deciles, chemical variables into quartiles, and age group used the ranges: (0–20), (20–30), (30–35),

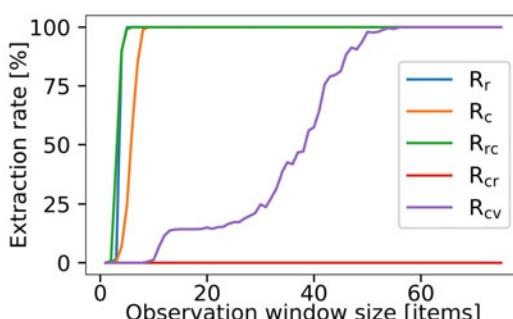


Fig. 3 Search results of observation window sizes for all atomic rule categories. LMC extracts all rules from the R_r , R_{rc} , and R_c categories when the observation window size is greater than the expected value of $p_c(t)$'s item timing. As the observation window grows, LMC extracts rules from R_{cv} , which are innocuous for the generating process separation task

(35–38), (38–40), (40–50), and (50–150). The poverty score was computed as the percentage of population per track that had a ratio between income and poverty level below 1.5. Additionally, deciles seven through ten were combined. Education score was calculated as the population percentage per track that had demonstrated education beyond high school level. Deciles eight through ten of the education score were merged. Finally, the race score was calculated as the percentage of non-White population per track.

Evaluation

Huang et al. [35] analysis provided a relevance criterion to interpret the extracted rules within their field. Therefore, the evaluation goal was to measure how many of Huang et al.’s criteria were extracted by the LMC miner.

For any newfound rule, the odds ratio (OR) was computed using a 95% confidence interval (CI).

The CI was calculated using 10000 runs of bootstrapping, following Huang et al. [35]’s evaluation of rule relevance. Table 1 lists the mined rules from Huang et al. [35] S → C scenario, with their respective support and lift. The LMC miner extracted only five rules from Huang et al.’s 13 rules.

Table 2 lists mined association rules in the S → S scenario. LMC found two out of the six rules reported by Huang et al. Additionally, LMC found three rules, highlighted in Table 2, that did not pass the min-support and lift criteria from Huang et al.

Table 3 shows the odds ratio (OR) and estimated 95% confidence interval (CI) for rules exclusively extracted by LMC in the S → S scenario. The OR analysis showed that the new rules had higher OR than the rules from Huang et al., whose OR ranged from 1.75–3.56. Rules that passed LMC had the largest OR in both mining scenarios. Thus, LMC rules are more likely to appear in a repeat experiment, and therefore LMC rules may be deemed more desirable.

Table 1 LMC’s extraction of socioeconomic and chemical exposure association rules (S → C). LMC mined five rules out of the 13 extracted using FRM. LMC rules had the highest lift

Rules	Support	Lift
Race score = 1 → Diesel = Q1	0.144	1.783
Race score = 1 → Butadiene = Q1	0.142	1.805
Race score = 1 → Toluene = Q1	0.138	1.746
Race score = 1 → Benzene = Q1	0.130	1.653
Race score = 1 → Acetaldehyde = Q1	0.126	1.596

Table 2 LMC extracted five rules in the S → S scenario. Two rules correspond to the FRM mined rules with highest lift, and three rules (in bold) did not pass Huang et al. [35] min-support threshold

Rules	Support	Lift
Age group = 40–50 → Race score = 1	0.172	1.585
Race score = 1 → Age group = 40–50	0.172	1.585
Poverty score = 1 → Education score = 8	0.038	3.919
Poverty score = 1 → Education score = 7	0.034	3.210
Race score = 1 → Age group = 50–150	0.015	2.227

Table 3 Odds ratio (OR) and confidence interval (CI) for LMC extracted rules in the S → S scenario. The 95% CI was estimated using 10,000 bootstrapping iterations

Rule	OR	Est. 95%	CI
Poverty score = 1 → Education score = 8	11.18	10.49	11.94
Poverty score = 1 → Education score = 7	6.74	6.35	7.17
Race score = 1 → Age group = 50–150	5.68	5.10	6.39

LMC rules provide meaningful insight, in particular on rarely, but consistently occurring relations, which may provide application experts new hypotheses to investigate. For example, as seen in the $S \rightarrow S$ scenario of the database mining examples, LMC provided additional rules that hint to a predominantly White ageing population (Race score = 1 → Age group 40–150), and to a correlation between low poverty score and high education levels (poverty score = 1 → education score = 7 or 8).

Rehabilitation Routine Mining

The following example illustrates how LMC can be used to interpret patient behavior during stays at a day care rehabilitation center. LMC and FRM miners try to classify patients into physically active and sedentary groups. However, results show that FRM represents the cohort's average behavior and thus fails to assign patients to groups.

Methodology

The rehabilitation routine mining examples used activity labels from the longitudinal stroke rehabilitation study of Derungs et al. [36]. The study was approved of the Swiss Cantonal Ethics Committee of the canton Aargau, Switzerland (Application number: 2013/009). There were 11 patients in the study, aged 34–75 years, among them five female and four used a wheelchair. In addition, data from a patient excluded from the original rehabilitation study [36] was added to the dataset, for a total of 12 patients.

Patients visited the day care center for approximately 3 days per week over 3 months to participate in individual and group training sessions, socialize with others, and follow personal activity preferences. Some training sessions available to patients were physiotherapy, ergotherapy, and training in the gym. Patients performed activities of daily living, including walking, eating and drinking, setting the table, writing, and making coffee. Behavior of each

patient was recorded for up to 8 h on 10 days at the center by two observers accompanying patients. In addition, body motion was recorded using inertial sensors attached to wrists, upper arms, and tight positions. During the observation time, the examiners annotated patient activities using a customized annotation tool on a smartphone, resulting in a total of 16,226 activity labels. Therapists scored patients for their ability to execute activities of daily living independently using the Extended Barthel Index (EBI) [37]. The EBI consists of 16 categories. Each category receives a score within the range zero to four, where zero means that the patient requires full support, and four means the patient can live independently.

Miners used the start of activity labels as timestamped items and a 20-minute observation window to create candidate rules by pairing the first item of the observation window with remaining items. The miners extracted atomic rules using their respective primary selection criterion. The same secondary criterion was used to filter the resulting rule sets and the remaining rules were assembled into graphs. Each resulting independent subgraph was considered a routine and a study observer assigned a label. The FRM miner's min-support threshold was set to 0.0038 with the goal of selecting the same number of rules as the LMC miner.

Evaluation

Miners were evaluated by submitting their extracted rules to the same post-processing two stage procedure: (1) secondary rule selection criterion and (2) graph-based routine classification. In the secondary rule selection stage, three methods were evaluated: Bayesian factor (Eq. 6), min-confidence threshold of 0.5, and best confidence per conclusion. With the retained rules a graph was constructed and routines extracted as independent subgraphs. For each mining method, the goal was to find the secondary rule selection criterion that provided a balance between activity label count per graph and the number of independent graphs.

A patient's contribution to the resulting routines was analyzed using a patient exclusion process (PEP), where a patient's data was removed from the dataset and resulting routines were compared with routines mined using all patients.

Based on the type of the majority of activities in the routine, a study observer named LMC routines as socializing, eating, using the phone, and intense and balance training. Whereas, FRM routines were named mobility, eating, and cognitive-motor training. Preliminary results showed that FRM routines lacked emphasis on activities related to socializing.

The *Primary Criteria* column in Fig. 4 shows the resulting graphs based on atomic rules extracted by LMC and FRM miners. Activities in both graphs are hyperconnected, i.e., multiple edges connect activities. However, for FRM, there are nodes with one edge. FRM rules do not describe the flow from one activity to another, but rather, the associations of repeating events, e.g., repetitions of an exercise. In contrast, LMC looks for successive activities, and the respective low count of activity transitions vs exercise repetitions does not affect the rule selection. For both mining algorithms, the hyperconnected graph yielded no useful routine information.

With a Bayesian factor ≥ 1 , LMC mined rules focus mostly on self-referencing activities, e.g., walking \rightarrow walking, resulting in single activity

subgraphs. In contrast, the Bayesian factor criterion removed most of the FRM rules. The resulting subgraphs had too few activities to consider them as routines. For FRM rules, the confidence secondary rule selection criterion reduced the graph size, but it was unable to create independent subgraphs. However, with LMC rules, the confidence secondary rule selection criterion selected many self-referencing rules creating two more subgraphs than the Bayesian factor, containing four activities each. Nevertheless, there were too many single activity subgraphs to consider the split as routines. The best balance was obtained between the number of subgraphs and activities per graphs using the best confidence per conclusion criterion. After secondary rule selection, LMC-mined rules yielded five routines, whereas FRM-mined rules yielded only three. Figure 4 illustrates the resulting subgraphs for each mining algorithm and secondary rule selection criteria.

Figure 5 illustrates an example of the changes in routine graphs when removing patients with active and sedentary behavior. For both mining algorithms, when removing one patient, the routine's activity composition varied, but the assignment of routine labels by the study observer did not vary. In the PEP analysis, using the best confidence per conclusion secondary criterion, LMC mined routines that grouped patients into

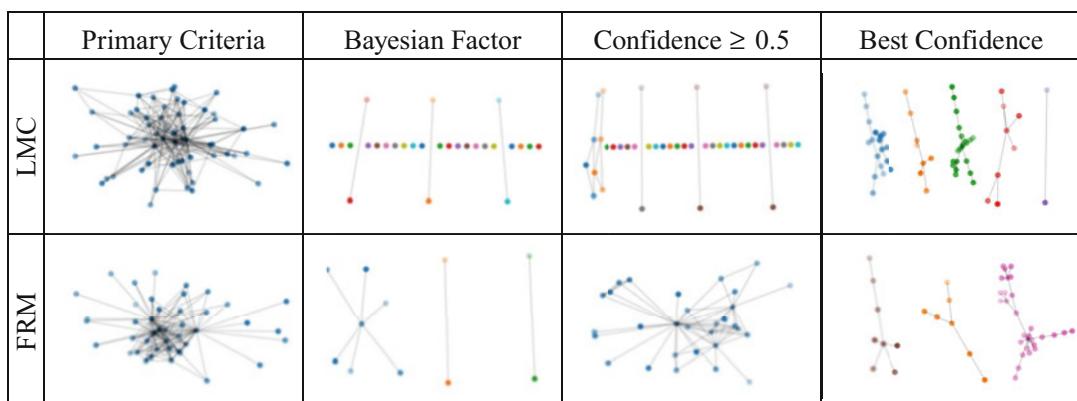


Fig. 4 Graphs constructed using rules derived by LMC, FRM, and different secondary rule selection criterion. Without a secondary rule selection criterion, both LMC and FRM produce a single graph and no useful routine

information is extracted. A balance between activity count per graph and number of independent graphs was achieved using the best confidence per conclusion secondary rule selection criterion

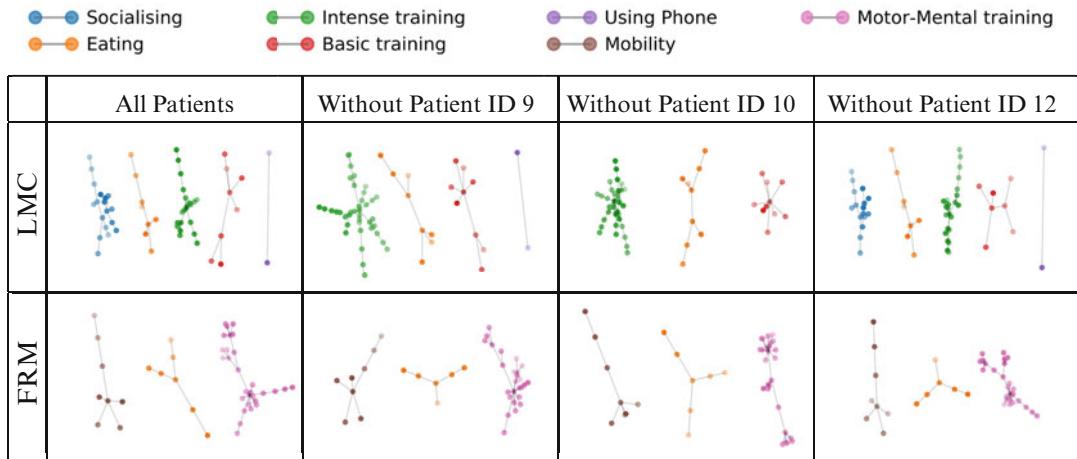


Fig. 5 Routine graphs when patients are removed from the dataset, i.e., PEP method. LMC's focus on rare rules highlights the importance of each individual patient's contribution to the graph representation. With PEP analysis,

physically active and sedentary groups, as indicated by the study observer. Removing patient ID 10 made the routine *using a phone* disappear. Apparently, patient 10 received calls while playing with the phone. The physically active group contained patient IDs 2,4,6,9, and 10. The physically active group refers to patients following their rehabilitation schedule closely. No relation to activity intensity or EBI score was found. The active patient group consisted of one wheelchair rider, patients with different EBI starting points, and some patients, where the EBI score did not change. When a patient from the active group was removed, the extracted routine number reduced to four and *socializing* was always missing. The result appears counterintuitive, as the sedentary group has been likely involved in socializing, but could be explained by the chosen 20-minute observation window, which causes LMC to focus on transitions between activities of at most 20 min duration. Sedentary patients would perform individual activities for periods longer than 20 min. Therefore, their socializing activities would not be associated into rules.

For comparison, PEP analysis for FRM-extracted routines using rules with the best confidence per conclusion found that the removal of any one patient did not affect the extracted

the LMC miner was able to separate patients into active and sedentary behavior groups. In contrast, the number of FRM-mined routines did not change under PEP analysis and no further insight was derived

routines. Therefore, FRM provided no further insight to classify patients into active or sedentary groups.

The rehabilitation routine mining examples illustrated the difference between both association rule mining criteria. Routines mined with FRM did not change during PEP analysis. FRM mined routines that were common to the entire population. With LMC, the routines changed during PEP analysis, grouping patients into active and sedentary groups. Hence, FRM answers the question: Which routines are common among patients?, and LMC answers the question: What types of patients are there?

Conclusions

FRM algorithms can be converted to use LMC by simply replacing the min-support threshold. As most algorithms exploit the support's closure property, and LMC has the same computational complexity as the calculation of support or confidence, there is no complexity penalty on any migrated algorithm. Albeit, the extracted rules will be different.

One limitation of LMC is the sporadic association of frequent premises with infrequent

conclusions into irrelevant rules. A secondary rule selection criterion can help remove irrelevant rules. However, the secondary rule selection criterion of choice depends on the application. For example, in the basic rule mining example a confidence threshold was used to separate items into generating processes, best rule confidence per conclusion worked for the rehabilitation routine mining task, and for the database mining example, no secondary rule selection criterion was needed.

Part of the motivation to introduce secondary rule selection criteria, other than confidence, is that association rules should provide some predictive power [15]. For LMC, rules have a predictive power of at least 50%. In other words, items in the premise of an LMC rule are colocated with items in the conclusion in at least 50% of the transactions. Although LMC was chosen to have better predictive performance than random chance for atomic rules, the 50% threshold used by LMC might be too restrictive. Lower threshold values might be necessary when considering categorical variables in the conclusion or conjunctive rules.

BRM, and in particular LMC, are not a replacement for FRM. The application should drive the choice of algorithm. For example, suppose a dataset contains symptoms, health-related behaviors, and disease outcomes. FRM is better suited to answer questions like “Which behaviours are most conducive to sickness?”. Whereas BRM is better suited to answer questions including “Which symptoms and behaviours correlate to a specific disease?”. The difference between ARM branches is summarized as follows: FRM focuses on extracting rules that describe commonalities between generating processes. In contrast, BRM looks for rules that describe each process.

This chapter showed that LMC does not suffer from support dilution, and that LMC is capable of extracting rare rules. The basic rule mining and socioeconomic examples illustrated how LMC extracted frequent and rare rules. In the rehabilitation routine mining examples, LMC was used to mine rules, create routines, and group patients into active and sedentary groups. Only LMC rules provided patient grouping information. As new medical datasets are investigated, LMC is a

powerful tool when considering ARM for knowledge extraction.

Acknowledgments The authors are thankful for the permission to utilize the datasets used for illustration in this chapter.

References

1. Agrawal R, Imielinski T, Swami A. Mining Association Rules Between Sets of Items in Large Databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data. SIGMOD '93. ACM; 1993. p. 207–216. Available from: <https://doi.org/10.1145/170035.170072>.
2. Lopera Gonzalez LI, Amft O. Mining Hierarchical Relations in Building Management Variables. *Pervasive and Mobile Computing*. 2016;26:91–101. Available from: <http://www.sciencedirect.com/science/article/pii/S1574119215001935>.
3. Liu S, Pan H. Rare itemsets mining algorithm based on RP-Tree and Spark framework. *AIP Conf Proc*. 1967 (1):040070. <https://doi.org/10.1063/1.5039144>.
4. Grabot B. Rule mining in maintenance: analysing large knowledge bases. *Comp Indust Eng*. 2018; 139:1–15. Available from: <https://hal.archives-ouvertes.fr/hal-02134705>
5. Li J, Fu AWc, Fahey P. Efficient discovery of risk patterns in medical data. 2009;45(1):77–89. Available from: <https://www.sciencedirect.com/science/article/pii/S0933365708000900>.
6. Bashir S, Jan Z, Baig AR. Fast algorithms for mining interesting frequent itemsets without minimum support. 2009, Available from: <http://arxiv.org/abs/0904.3319>.
7. Djenouri Y, Djenouri D, Belhadi A, Fournier-Viger P, Lin JCW. A new framework for metaheuristic-based frequent itemset mining. *Appl Intell*. 2018;48 (12):4775–4791. Available from: <https://doi.org/10.1007/s10489-018-1245-8>.
8. Tahyudin I, Nambo H. The combination of evolutionary algorithm method for numerical association rule mining optimization. In: Xu J, Hajiyev A, Nickel S, Gen M, editors. *Proceedings of the tenth international conference on management science and engineering management. Advances in intelligent systems and computing*. Singapore: Springer. 2017;p. 13–23.
9. Borah A, Nath B. Identifying risk factors for adverse diseases using dynamic Rare association rule mining. *Expert Syst Appl*. 2018;113:233–263. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0957417418304251>.
10. Li J, Fu AWc, He H, Chen J, Jin H, McAullay D, et al. Mining risk patterns in medical data. In: *Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining. KDD '05*.

- ACM; 2005. p. 770–775. Available from: <https://doi.org/10.1145/1081870.1081971>.
11. Erwin A, Gopalan RP, Achuthan NR. Efficient mining of high utility itemsets from large datasets. In: Advances in knowledge discovery and data mining. Springer, Berlin, Heidelberg; 2008. p. 554–561. Available from: https://doi.org/10.1007/978-3-540-68125-0_50.
 12. Fournier-Viger P, Lin JCW, Truong-Chi T, Nkambou R. A survey of high utility itemset mining. In: High-utility pattern mining. Cham: Springer; 2019. p. 1–45. https://doi.org/10.1007/978-3-030-04921-8_1.
 13. Nguyen LTT, Mai T, Vo B. High utility association rule mining. In: High-utility pattern mining. Cham: Springer; 2019. p. 161–74. https://doi.org/10.1007/978-3-030-04921-8_6.
 14. Zaki M. Scalable algorithms for association mining. IEEE Trans Knowl Data Eng. 2000;12(3):372–90.
 15. Lin WY, Tseng MC, Su JH. A confidence-lift support specification for interesting associations mining. In: Chen MS, Yu PS, Liu B, editors. Advances in knowledge discovery and data mining, Lecture notes in computer science. Berlin: Springer; 2002. p. 148–58.
 16. Brin S, Motwani R, Silverstein C. Beyond market baskets: generalizing association rules to correlations. In: Proceedings of the 1997 ACM SIGMOD international conference on management of data. SIGMOD '97. ACM; 1997. p. 265–276. <https://doi.org/10.1145/253260.253327>.
 17. Brin S, Motwani R, Ullman JD, Tsur S. Dynamic itemset counting and implication rules for market basket data. In: Proceedings of the 1997 ACM SIGMOD international conference on management of data. SIGMOD '97. ACM; 1997. p. 255–264. <https://doi.org/10.1145/253260.253325>.
 18. Yan X, Zhang C, Zhang S. Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support. Expert Syst Appl. 2009;36(2):3066–76.
 19. Liu L, Wang S, Peng Y, Huang Z, Liu M, Hu B. Mining intricate temporal rules for recognizing complex activities of daily living under uncertainty. Pattern Recogn. 2016;60:1015–28. Available from: <http://www.sciencedirect.com/science/article/pii/S003132031630173X>
 20. Srinivasan V, Koehler C, Jin H. RuleSelector: selecting conditional action rules from user behavior patterns. Proc ACM Interact Mobile Wearable Ubiquitous Technol. 2018;2(1):35:1–35:34. <https://doi.org/10.1145/3191767>.
 21. Padillo F, Luna JM, Herrera F, Ventura S. Mining association rules on big data through mapreduce genetic programming. Integr Comp Aided Eng. 2017;25(1):31–48. <https://doi.org/10.3233/ICA-170555>.
 22. Guillame-Bert M, Crowley JL. Learning temporal association rules on symbolic time sequences. In: Proceedings of the 4th Asian conference on machine learning, ACML; 2012. p. 159–174.
 23. Liu B, Hsu W, Ma Y. Mining association rules with multiple supports. In: Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining. KDD '99. ACM; 1999. p. 15–18.
 24. Tsang S, Koh YS, Dobbie G. RP-Tree: rare pattern tree mining. In: Data warehousing and knowledge discovery. Berlin, Heidelberg: Springer; 2011. p. 277–88. https://doi.org/10.1007/978-3-642-23544-3_21.
 25. Webb GI. OPUS: an efficient admissible algorithm for unordered search. J Artif Intell Res. 1995;3:431–65. Available from: <https://www.jair.org/index.php/jair/article/view/10152>
 26. Webb GI. Efficient search for association rules. In: Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining. KDD '00. ACM; 2000. p. 99–107. Available from: <https://doi.org/10.1145/347090.347112>.
 27. Fournier-Viger P, Tseng VS. Mining top-K NoN-REDUNDant association rules. In: Chen L, Felfernig A, Liu J, Raš ZW, editors. Foundations of intelligent systems, Lecture notes in computer science. Berlin: Springer; 2012. p. 31–40.
 28. Cheung DW, Han J, Ng VT, Wong CY. Maintenance of discovered association rules in large databases: an incremental updating technique. In: Proceedings of the twelfth international conference on data engineering; 1996. p. 106–114.
 29. Tobji MB, Gouider M. Incremental maintenance of association rules under support threshold change. In: Proceedings of the IADIS international conference on applied computing. IADIS; 2006. Available from: <http://arxiv.org/abs/1701.08191>.
 30. Aqra I, Abdul Ghani N, Maple C, Machado J, Sohrabi SN. Incremental algorithm for association rule mining under dynamic threshold. Appl Sci. 2019;9(24):5398. Available from: <https://www.mdpi.com/2076-3417/9/24/5398>
 31. Tian D, Gledson A, Antoniades A, Aristodimou A, Dimitrios N, Sahay R, et al. A Bayesian association rule mining algorithm. In: 2013 IEEE international conference on systems, man, and cybernetics. IEEE; 2013. p. 3258–3264.
 32. Gay D, Boullé M. A Bayesian approach for classification rule mining in quantitative databases. In: Machine learning and knowledge discovery in databases. Berlin, Heidelberg: Springer; 2012. p. 243–59. https://doi.org/10.1007/978-3-642-33486-3_16.
 33. Lopera Gonzalez LI. Mining functional and structural relationships of context variables in smart-buildings [PhD Thesis]. 2018. Available from: <https://opus4.kobv.de/opus4-uni-passau/frontdoor/index/index/docId/573>.
 34. Lopera Gonzalez LI, Derungs A, Amft O. A Bayesian approach to rule mining. 2019. Available from: <https://arxiv.org/abs/1912.06432v1>.
 35. Huang H, Tornero-Velez R, Barzyk TM. Associations between socio-demographic characteristics and chemical Concentrations contributing to cumulative

- exposures in the United States. *J Expos Sci Environ Epidemiol.* 2017;27(6):544–50. <https://doi.org/10.1038/jes.2017.15>.
36. Derungs A, Schuster-Amft C, Amft O. Longitudinal walking analysis in hemiparetic patients using wearable motion sensors: is there convergence between body sides?. *Front Bioeng Biotechnol.* 2018;6. <https://doi.org/10.3389/fbioe.2018.00057/full>.
37. Prosiegel M, Böttger S, Schenk T, König N, Marolf M, Vaney C, et al. Der Erweiterte Barthel-Index (EBI)—eine Neue Skala Zur Erfassung von Fähigkeitsstörungen Bei Neurologischen Patienten. *Neurol Rehabil.* 1996;1:7–13.



Joseph Davids, Kyle Lam, Amr Nimer, Stamatia Gianarrou, and Hutan Ashrafiyan

Contents

Introduction	320
AI in Various Specialty Delivered Medical Education	326
Literature Review	326

J. Davids (✉)

Imperial College London NHS Trust, London, UK

Hamlyn Centre for Robotics and Artificial Intelligence,
Department of Surgery and Cancer, Imperial College
London, London, UK

National Hospital for Neurology and Neurosurgery Queen
Square, London, UK

e-mail: j.davids@imperial.ac.uk

K. Lam

Hamlyn Centre for Robotics and Artificial Intelligence,
Department of Surgery and Cancer, Imperial College
London, London, UK

Imperial College London NHS Trust, London, UK

A. Nimer

Imperial College London NHS Trust, London, UK

S. Gianarrou

Hamlyn Centre for Robotic Surgery and AI, Department of
Surgery and Cancer, Imperial College London, London,
UK

H. Ashrafiyan

Department of Surgery and Cancer, Imperial College
London NHS Trust, London, UK

Institute of Global Health Innovation, Imperial College
London, London, UK

Hamlyn Centre for Robotics and Artificial Intelligence,
Department of Surgery and Cancer, Imperial College
London, London, UK

Discussion	328
Meta-Analysis	328
Conclusion	336
References	336

Abstract

Artificial intelligence (AI) is making a global impact on various professions ranging from commerce to healthcare. This section looks at how it is beginning and will continue to impact other areas such as medical education. The multifaceted yet socrato-didactic methods of education need to evolve to cater for the twenty-first-century medical educator and trainee. Advances in machine learning and artificial intelligence are paving the way to new discoveries in medical education delivery.

Methods

This chapter begins by introducing the broad concepts of AI that are relevant to medical education and then addresses some of the emerging technologies employed to directly cater for aspects of medical education methodology and innovations to streamline education delivery, education assessments, and education policy. It then builds on this to further explore the nature of new artificial intelligence concepts for medical education delivery, educational assessments, and clinical education research discovery in a PRISMA-guided systematic review and meta-analysis.

Results

Results from the meta-analysis showed improvement from using either AI alone or with conventional education methods compared to conventional methods alone. A significant pooled weighted mean difference ES estimate of ES 4.789; CI 1.9–7.67; $p = 0.001$, $I^2 = 93\%$ suggests a 479% learner improvement across domains of accuracy, sensitivity to performing educational tasks, and specificity. Significant amount of bias between studies was identified and a model to reduce bias is proposed.

Conclusion

AI in medical education shows considerable promise in domains of improving learners' outcomes; this chapter rounds off its discussion with the role of AI in simulation

methodologies and performance assessments for medical education, highlighting areas where it could augment how we deliver training.

Keywords

Artificial intelligence · Medical education · AI in surgical education · Performance assessment · AI in dermatology education · AI in ophthalmology education · AI in surgical education · AI in simulation · Deep learning · Neural networks · Machine learning · Meta-analysis

Introduction

The origins of modern-day medical education can be traced from Alexandria in Egypt, dating back to 3000 BC, to Ancient Greece and the works of Hippocrates and Greco-Roman doctors like Galen of Pergamon who developed an understanding of human and animal anatomy and would encourage his students to examine the bodies of those who had perished in gladiatorial combat. Fulton provides a more detailed account of the history of medical education [1]. In many ways, medical education has not changed very much over the centuries; however, the demands and necessities of modern life make it increasingly apparent that technology will play an ever more important role.

Similarly, concepts of thinking machines have centuries of history, from Descartes discourses in 1637 to Babbage's analytical engines [2]. The automaton was also explored well before antiquity, with biblical references to the times of King Solomon. But according to the Royal Society's archives, Jacques de Vaucanson, an eighteenth-century mechanician from Grenoble, devoted much of his attention to the exploration of

automata, unveiling public depictions of these autonomous machines [3, 4]. These depictions sparked considerable public interest, rhetoric and criticism about mechanical creations. That Voltaire credited Vaucanson as a rival of Prometheus “in his ability to steal fires in his search to give life” encouraged and illustrated the need to educate the public about the origins of machine learning – the medical community is no different needing a shift in our thought process and approach to AI and its applications [5]. From Friar and Marshall’s assembled oeuvre on the motives of de Vaucanson, one particular quote highlights the importance of this viewpoint [6].

...One such motive might be the desire to understand how some part of the natural world actually works. Derek de Solla Price [5] and Silvio Bedini [6] have recently argued that many early gadgets and machines were neither ‘trivial toys’ nor ‘immediately useful inventions.’ Rather, they were simulacra. That is, the devices were models of a very special sort, models ‘... whose very existence offered tangible proof, more impressive than any theory, that the natural universe of physics and biology was susceptible to mechanistic explication.’ [7] This leads Price to reverse the usual interpretation of the relationship between high technology and ‘pure science’ in the Hellenistic and Roman world. It is not the case, he suggests, that ‘... certain theories in astronomy and biology derived from man’s familiarity with various machines and mechanical devices.’ On the contrary, ‘... some strong innate urge toward mechanistic explanation led to the making of automata, and ... from automata has evolved much of our technology ...’ [8]. Another alternative might, of course, be that this urge toward mechanistic explanation led both to the construction of automata and to our practical technology....

In the 1950s, Alan Turing posed the question of whether machines could think, fundamentally forming the Turing test. This test involves a human interaction with a machine, and if an assessor human cannot differentiate between the human and the machine, then the machine is said to be thinking. Ashrafian et al. modified this test in 2014 for more medical implications for machine intelligence [7]. AlphaGo by DeepMind has shown that artificially intelligent agents can learn to become highly adept at certain tasks and show superior intelligence in domain-specific facets that previously were deemed beyond their

intelligence [8, 9]. Failures in how an AI is trained have also surfaced in modern times with one example being the Microsoft Tay, an AI agent that was designed to interact with Twitter users, but which unfortunately inherited chauvinistic tendencies based on the approach used in its training [10]. AI has considerable potential, but like any child starting to learn, it will require the correct type of data and the right guidance to become effective for it to be useful in education. Training when the premise of learning is flawed becomes suboptimal, and so too will the eventual outcome and educational approach. Education here has a twofold self-fulfilling philosophy: training the AI to learn how to educate the medical community and the medical community themselves learning about how the AI learns to teach it to learn optimally.

As this chapter was written during the global COVID-19 pandemic, our current approach to delivering education must evolve to keep up with twenty-first-century lifestyles. The imposed restrictions have also brought into focus the need to recalibrate improvements in work-life balance. However, the introduction of working-time directives has also affected training time and reduced the time to gain the required clinical experience to be deemed an expert in one’s specialty of interest. Therefore, there is an argument for leveraging other approaches to supplement the lack of experience by evolving how training is delivered using a trans-specialty and multidisciplinary approach. AI is a candidate that can support this notion of personalized medical education through agent-based modeling and similar paradigms.

The recent debates surrounding machine learning and artificial intelligence can also be extended to include medical education for clinicians in low- and middle-income countries, thus opening a synergistic global education platform to a wider audience from different cultural demographics. There is also the opportunity for the joint education of members within the clinical multidisciplinary team. By learning together, the team members can foster more of an appreciation for how each member’s role impacts the team-working dynamics and the individual clinical challenges that each must overcome for optimal patient care. This is

paramount from a human factors' perspective not just for effective learning delivery to an ever-changing workforce, but also to safeguard our patients and ensure better outcomes.

Fictional anthropomorphisms for AI have mainly dominated public discussions about their impact on humanity, altering perspectives, be it in a positive or negative way. However, little has been explored for the fact that learning about and understanding the human body and interactions with AI may have aspects where anthropomorphic AI applications become warranted. A prime example being areas related to *in situ* medical simulation where an AI agent may be useful to automate a specific task to improve fidelity and timing and enable fine adjustments to be made to help guide the trainee to the optimal management approach for a task.

Another area that is worth considering, which adds a challenge to AI for medical education, pertains to the various learning style models that exist in education theory. David Kolb's model of experiential learning states that some individuals learn continuously in a process that strengthens specific areas and weakness in our knowledge acquisition. This gives rise to four areas of preferential learning styles: 1) accommodating, 2) converging, 3) diverging, and 4) assimilating. Accommodators prefer a hands-on approach to experiential learning; convergers adopt and test out problems through abstractions to help them better learn; divergers prefer using their own personal experiences to develop widely applicable theoretical frameworks to guide their learning; and assimilators like convergers develop abstract concepts into their own original theories [11–13].

Similarly, Honey and Mumford evolved Kolb's model to four main learning styles – 1) activists, 2) pragmatists, 3) reflectors, and 4) theorists [14, 15] – clearly elucidating that individuals might fall within one or more of these categories. Others such as Gregorc and Butler's mind style model looks into how we think or frame our thought processes, categorizing us into a spectrum of either 1) concrete sensate thinkers or 2) abstract idea generators versus either 3) sequential logic-based linear thinkers or 4) unpredictable

and multidirectional random thinkers [16]. Adding another dimension to the challenge of classifying learning styles, Barbe and colleagues' VARK model highlights important differences between the approaches for 1) visual, 2) auditory, and 3) kinesthetic learning styles [17]. Although some of these models have consistent similarities, each applies to various individual learners, which lends itself to considering a personalized approach to medical education.

Thus, it is unsurprising that our current methods of medical education remain ill-equipped to cater for all these types of learners. Often the status quo attempts to place a learner into a particular style, perhaps due to inherent resource constraints or because the situation being dealt with dictates a particular approach be adopted that may not be suitable to that individual or the organization for that matter. Nevertheless, the best educators and educational systems can identify the needs of the learner and what type of learning style suits them best in a cost-effective way. This is where AI algorithms could potentially augment medical education, and this remains an area of active research [18] as Table 1 summarizes. Identifying the specific needs of the learner and providing decision support that allows them to be effective at what they are learning could lead to a productive workforce with complementary skill sets.

Another point worth mentioning is that most practitioners have genuine concerns about whether AI algorithms will replace them as educators or as clinicians, while others pragmatically question whether an algorithm can be as good as a human medical teacher. We attempt to dispel some of the myths and allay some fears while keeping the conversation open so that both proponents and opposing ideas can flourish, thus emboldening thought and stimulating the discourse that will move the field forward.

In this section, we explore how AI is being used in medical education, its advantages, and pitfalls and where innovations have been made or could evolve to enhance this for medical practitioners. This will not be in the form of the usual rhetoric of dystopian medical universes but

Table 1 The AI models that have been discussed and useful buzzwords that nonengineering medical communities need to be familiar with in the twenty-first-century artificial intelligence for medical education

Author and year	Countries involved	Study group	AI models and platforms discussed and interesting AI buzzwords	Subspecialty discussed
Bastardot et al. 2019 [20]	Switzerland	Undergraduates, postgraduates	Deep learning	AI for radiology image and in cardiology for ECG interpretation
van der Niet and Bleakley 2020 [21]	United Kingdom and Netherlands	–	Artificial neural network (ANN), actor network theory (ANT), object-oriented ontology, INTERNIST I, MYCIN, CADUCEUS	General Calculated Medical Pedagogy and Genomics CRISPR
Lindqwister et al. 2020 [22]	United States	Residents	K-nearest neighbor (KNN)	Radiology curriculum design, preparation, AI topic interest, use of didactic methods, Journal Clubs, pacing and assessment
Carin 2020 [23]	United States	–	Deep learning	Radiology and image and video analysis in radiology, ophthalmology, and dermatology; deep learning as a tool for quantitative assessment
Clancey 1983 [24]	United States	Students	ATTENDING, GUIDON, MYCIN consultation system, Intelligent Computer- Aided Instructional Systems (ICAI)	GUIDON uses mixed initiative dialogue and the MYCIN infectious disease diagnoses framework developed as a computer-based instruction platform for teaching diagnosis
Bourlas et al. 1995 [25]	Greece	Undergraduates, postgraduates, continuing medical education	CARDIO-LOGOS platform, reference model and AI techniques	AI leveraging the page-turning architecture for cardiology ECG pattern recognition
Voss et al. 2000 [26]	Germany	Undergraduates, postgraduates	LAHYSTOTRAIN, Virtual Reality and Intelligent Tutoring System	Artificially intelligent laparoscopic and hysteroscopy trainer and surgical simulator, general surgery
Stasui et al. 2001 [27]	United States	Students	CARDIOLOG	Artificially intelligent echocardiography tutor for cardiology
Kintsch 2002 [28]	United States	Second- to 4th-year medical undergraduates	Latent Semantic Analysis, Singular Value Decomposition, Dimensionality Reduction, Intelligent Essay Assessor	Clinical case summaries using a semantic space construction approach, for history of present illness, physical examination, laboratory data, differential diagnosis in general medical education and clinical psychology
Caudell et al. 2003 [29]	United States	Undergraduates	Touch Simulation Engine, National Computational Science Alliance AG distributed networking	An AI-powered platform for virtual reality simulation and problem-based general medical education

(continued)

Table 1 (continued)

Author and year	Countries involved	Study group	AI models and platforms discussed and interesting AI buzzwords	Subspecialty discussed
Crowley and Medvedeva 2003 [30]	United States	Undergraduates	SlideTutor an intelligent tutoring system, domain model ontology, domain task ontology, Dynamic Solution Graph	AI in diagnostic education for dermatology and inflammatory skin diseases
Michael et al. 2003 [31]	United States	First-year undergraduate medical students	CIRCSIM-Tutor, Using Natural Language Processing	Cardiovascular medicine and physiology on the baroreceptor reflex
McFadden and Crim 2016 [32]	United States	Continuing medical education	Knowledge- Based Intelligent tutor (KBIT), convenience sampling methodology	Web-delivered multimedia-based training, family medicine and primary care, general medical case vignettes, rheumatology education
Khumrin et al. 2017 [33]	Australia	Medical undergraduates	Naïve Bayes, support-vector machine (SVM), Artificial Neural Networks (ANN), Zero R, J48 decision tree, and Logitboost (using DecisionStump as a classifier)	Gastroenterology AI e-learning platform using preprocessed clinical features and machine learning models
Chen et al. 2019 [34]	United States	Postgraduate radiology residents	Trove Dashboard, Convolutional Neural Networks and RELU, LSTM, Natural Language Processing, tf-IDF, GRU, SMART stoplists, latent dirichlet allocation (LDA), Word2Vec and GloVe (Word embedding), latent semantic indexing (LSI), probabilistic latent semantic indexing (pLSI), Bag of Words	Radiological keyword identification and mapping of diagnosis to trove and ICD codes, e.g., lymphangiomyomatosis
Cheng et al. 2020 [35]	Taiwan	Randomized controlled trial on undergraduate medical students	The AI education system – HipGuide, deep learning algorithm of DenseNet-121, grad-CAM algorithm, AI-augmented X-rays	AI in orthopedics diagnostics and interpretation of a hip fracture, analysis of pelvis anteroposterior (AP) view radiographs (PXRs)
ElSaadawi et al. 2009 [36]	United States	Pathology residents	ReportTutor, Natural Language Interface, cognitive intelligent tutoring systems	Intelligent Tutoring System (ITS) in diagnostic dermatopathology, a multisystem interface platform
Chieu et al. 2010 [37]	United States and France	Postgraduate student	TELEOS project, Temporal Bayesian networks for student modeling	Orthopedic screw fixation education using intelligent tutoring system for fine-grained didactic analysis
Fernández-Alemán et al. 2016 [38]	Spain	Medical students	Audience response system iSIDRA(Sistema De Respuesta Inmediata de la Audiencia in Spanish), Deep Neural Network	Locomotor system education, general and descriptive anatomy using acquisition and retention methods like GAHMS
Paranjape et al. 2019 [39]	Netherlands and Singapore	–	Natural Language Processing, Deep Learning, virtual agents, black-box model, and AI	Discussion of AI across all medical specialties and the challenges faced

(continued)

Table 1 (continued)

Author and year	Countries involved	Study group	AI models and platforms discussed and interesting AI buzzwords	Subspecialty discussed
Lang and Repp 2020 [40]	Germany	Students	NeuroTronics, with parallels from biology and electronics	Multispecialty education in the form of seminars, guided discussions, and demonstration and concrete development of algorithms. The content and the organization of the events were coordinated by the SciTecMed/NWTmed projects
Frize and Frasson 2000 [41]	Canada	–	Decision-support tools, such as scoring systems, Bayesian models, neural networks, to cognitive models	–
Masters 2019 [42]	Oman	Continuing medical education	AlphaGo, Psychological Reorientation	Teaching across all levels of medicine including continuing medical education and how it is going to continue to evolve
Holden et al. 2019 [43]	Canada	Postgraduate trainees and experts	Perk Tutor Decision tree and fuzzy rule-based assessment, Naïve Bayes, rule-based learners, zero rule regression, support vector regression, linear regression	A configurable assessment method that performs comparably to state-of-the-art methods and provides useful feedback for training
Mirchi et al. 2020 [44]	Canada	Postgraduate trainees and experts/fellows	Explainable AI platform for simulation education, single processing unit known as the perceptron, linear support-vector machine (SVM), virtual operative assistant	Neurosurgery, virtual operative assistant used for objective structured feedback of performance and skills education for a virtual reality brain tumor resection task
Bissonnette et al. 2019 [45]	Canada	Postgraduate trainees seniors and juniors and medical students	Linear discriminant analysis, k-nearest neighbors, naive Bayes, decision tree, support-vector machine, NeuroVR neurosurgical simulator, leave-one-out cross-validation	Spinal orthopedic and neurosurgery, virtual reality hemilaminectomy task skill discrimination between different levels of trainees. Surgical performance and novel metric analysis
Winkler-Swartz et al. 2019 [46]	Canada and Iran	–	Machine Learning to Assess Surgical Expertise (MLASE) checklist, Supervised and Unsupervised Machine Learning, hidden Markov models, support-vector machines, and artificial neural networks	Use of machine learning to assess surgical expertise in virtual reality simulation; it educates and allows a standardized and robust unbiased reporting metric for AI in publications across medical fields
Wang and Majewicz 2018 [47]	United States	Surgeons, novice to experts	Deep learning, convolutional neural networks, kNN (K-nearest neighbor, logistic regression (LR), novel vector space model (VSM), linear discriminant analysis (LDA), support-vector machine, nearest neighbor (NN), novel feature fusion, hidden Markov model, JIGSAWS, multinomial cross-entropy cost, Xavier initialization method, mini-batch updates of gradient descent	

practical applications of AI in medical education. We begin the chapter by presenting a recent systematic review of the literature looking at AI in medical education and discuss specialty areas where influences in AI for medical education are showing promise.

AI in Various Specialty Delivered Medical Education

Literature Review

To identify work that has already been done in AI for medical education, a systematic search of the literature was performed using the terms “artificial intelligence” and “medical education” in the following databases: Books@Ovid November 30, 2020; Journals@Ovid Full Text December 07, 2020; Your Journals@Ovid, AMED (Allied and Complementary Medicine) 1985 to November 2020; Embase Classic+Embase 1947 to 2020 December 07; Ovid Emcare 1995 to 2020 Week 49; Global Health Archive 1910 to 1972; Health and Psychosocial Instruments 1985 to October 2020; HMIC Health Management Information Consortium 1979 to November 2020; Ovid MEDLINE(R) ALL 1946 to December 04, 2020; Maternity and Infant Care Database (MIDIRS) 1971 to November 2020; and PUBMED.

Inclusion and Exclusion Criteria

Primary articles and reviews that made no specific reference to artificial intelligence and medical education were excluded. Included articles strictly discussed artificial intelligence and medical education, referring to them in their title and abstract. Also considered for inclusion were intelligent tutoring systems, but with a discussion on artificial intelligence in medical and surgical education.

Meta-Analysis

A meta-analysis was performed using R (Version 1.3.1093, RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA) and Stata (Stata Corp. 2017. Stata Statistical Software: Release 15. College Station,

TX: StataCorp LLC). A random effects model was implemented with I^2 to characterize the degree of heterogeneity. Ratio of means was calculated using a standardized difference approach as $(Me - Mc)/Mc$. There was no imputation of missing values. A p-value <0.05 was deemed statistically significant.

Bias Assessment

Bias was assessed using the Cochrane risk-of-bias tool.

Results

The search results identified 3102 publications, screened by KL and JD on the Rayyan platform with AN/HA acting as the tie breaker when necessary [19]. A PRISMA flowchart summarizes the results of the systematic review as shown in Fig. 1.

Results from Screening

Figure 1 shows the PRISMA flowchart for the results of the study. Studies were reported between 2000 and 2020. Twenty-eight studies were reviewed in the qualitative synthesis, and three studies were included in the final meta-analysis with one study presenting incomplete data on its standard deviations.

Bias Assessment Result

A significant risk of bias was seen across all the studies selected for the final meta-analysis (Table 2).

Meta-Analysis

Figures 2, 3, and 4 summarize the meta-analysis performed over 3 main studies with 14 sub-studies reporting on improvement over various domains. Majority of the sub-studies were extracted from Cheng et al. 2020 and looked at facets such as the improvement from leveraging AI for pre- and post-learning sensitivity, pre- and post-learning specificity, and pre- and post-learning accuracy. Cheng also looked at gained sensitivity, specificity, and accuracy, whereas Michael and colleagues explored the improvement in knowledge and performance between pre- and post-test scores. Aleman-Fernandez et al. studied an AI-powered

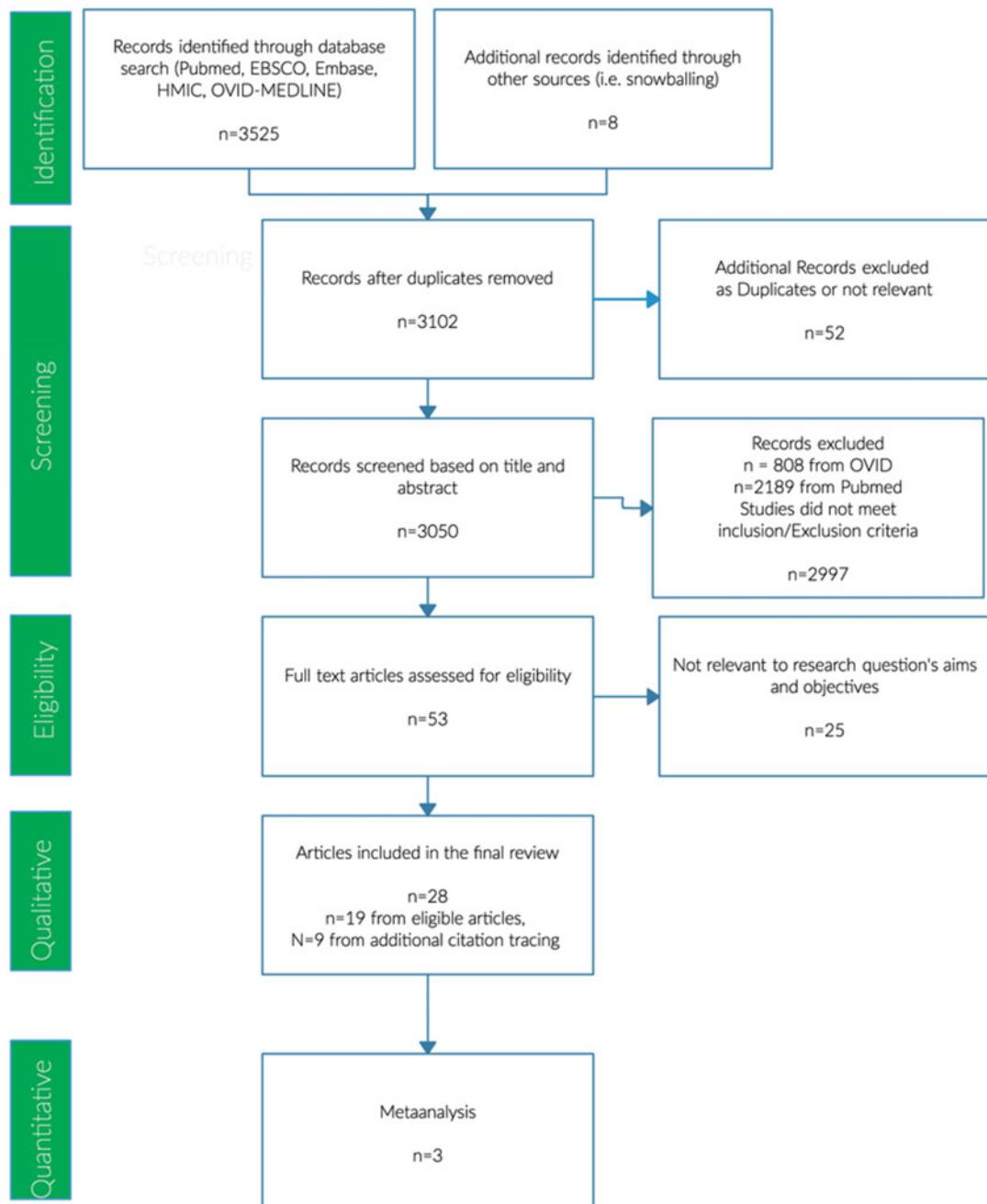


Fig. 1 PRISMA flowchart illustrating the number of articles selected for meta-analysis

intelligent tutoring system compared with a conventional approach.

Improvement from either using AI alone or with conventional education methods compared to conventional methods alone showed 1) a

significant pooled weighted mean difference (WMD) effect size (ES) estimate of ES 478.9%; CI 1.9–7.67; $p = 0.01$, $I^2 = 93\%$; 2) a significant pooled standardized mean difference (SMD) ES estimate of ES 87.8%; CI 0.352–1.404; $p = 0.001$,

Table 2 Cochrane risk-of-bias grid illustrating a high percentage of bias across some of the domains of blinding and allocation. [31, 35, 38].

Table 2 Cochrane Risk of Bias grid	Fernández-Alemán et al 2016	Michael et al 2003	Cheng et al 2020	Final Score
Random sequence generation (selection bias)	High Risk 😕	High Risk 😕	Low risk 😊🏃	66.6% 😕
Allocation concealment (selection bias)	High Risk 😕	High Risk 😕	High Risk 😕	100% 😕
Blinding of outcome assessment (detection bias) (patient-reported outcomes)	High Risk 😕	High Risk 😕	High Risk 😕	100% 😕
Blinding of participants and personnel (performance bias)	High Risk 😕	High Risk 😕	High Risk 😕	100% 😕
Blinding of outcome assessment (detection bias) (Mortality)				
Incomplete outcome data addressed (attrition bias) (Short-term outcomes (2-6 weeks))	low risk 😊🏃	High Risk 😕	low risk 😊🏃	66% 😊🏃
Incomplete outcome data addressed (attrition bias) (Longer-term outcomes (>6 weeks))	low risk 😊🏃	High Risk 😕	low risk 😊🏃	66% 😊🏃
Selective reporting (reporting bias)	low risk 😊🏃	high Risk 😕	low risk 😊🏃	66% 😊🏃

$I^2 = 85\%$; and 3) a significant pooled ratio of means (ROM) ES estimate of ES 94.1%; 0.937–0.944; $p = 0.001$, $I^2 = 100\%$.

Discussion

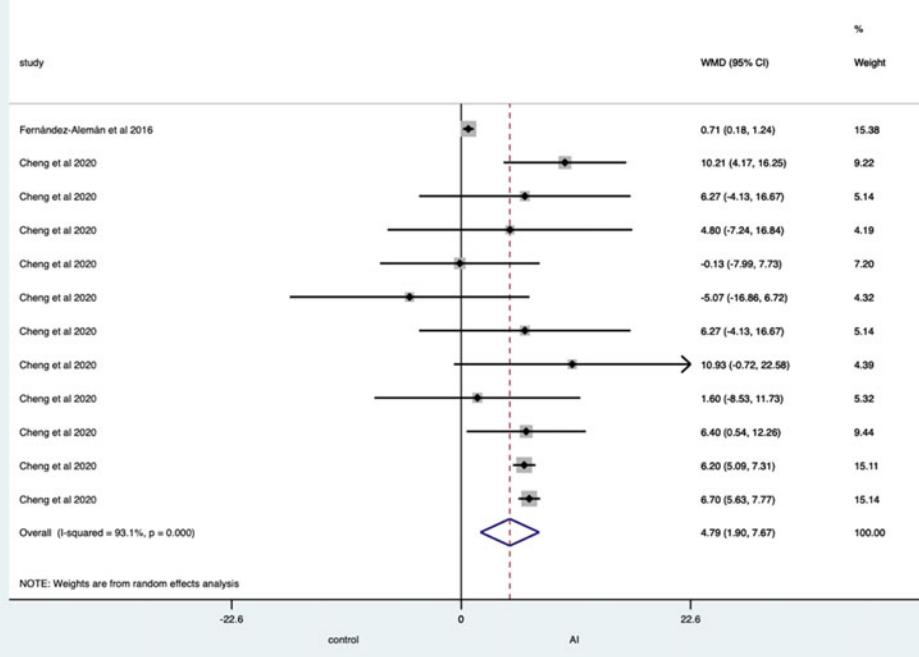
There is sometimes a separatist mentality in the academia of one field being pitched against the other (engineering against medicine with a never shall the two cultures mix attitude), but the boundaries between specialties are starting to blur with the applications of artificial intelligence within medicine facilitating crucial translationally relevant multidisciplinary collaborations. The puritanical practitioners of the latter specialty have on many occasions felt there is no reason to learn a bit about the former and vice versa. Members from the former camp that look at artificial intelligence as purely an engineering and technically demanding concept that only belong to engineers, fail to appreciate that subsets of AI such as deep neural networks derived its inspiration from the latter medical camp. So developing artificial

intelligence was born out of a clinical need to tackle and understand how the human brain works by leveraging either specialties to advance our knowledge, thus creating a cyclical irony of sorts. This chapter considers the interplay between these two specialties to be quintessential to translational innovation that will improve clinician education with the ultimate outcome being improved patient care. And perhaps AI can serve as the glue that unites specialties to answer some of the essential medical education questions and learning theory. Here we performed a very recent addition to the myriad of past systematic reviews that have demonstrated the growing adoption of AI in various clinical subspecialties from the context of medical education.

Meta-Analysis

Our meta-analysis demonstrates that the use of artificial intelligence significantly augments medical education in various domains of improvement, including improved learning accuracy,

WMD



WMD

Study	WMD	[95% Conf. Interval]	% Weight
Fernández-Alemán e	0.711	0.180 1.242	15.38
Cheng et al 2020	10.210	4.173 16.247	9.22
Cheng et al 2020	6.270	-4.129 16.669	5.14
Cheng et al 2020	4.800	-7.236 16.836	4.19
Cheng et al 2020	-0.130	-7.986 7.726	7.20
Cheng et al 2020	-5.070	-16.860 6.720	4.32
Cheng et al 2020	6.270	-4.129 16.669	5.14
Cheng et al 2020	10.930	-0.716 22.576	4.39
Cheng et al 2020	1.600	-8.527 11.727	5.32
Cheng et al 2020	6.400	0.540 12.260	9.44
Cheng et al 2020	6.200	5.086 7.314	15.11
Cheng et al 2020	6.700	5.632 7.768	15.14
D+L pooled WMD	4.789	1.904 7.674	100.00

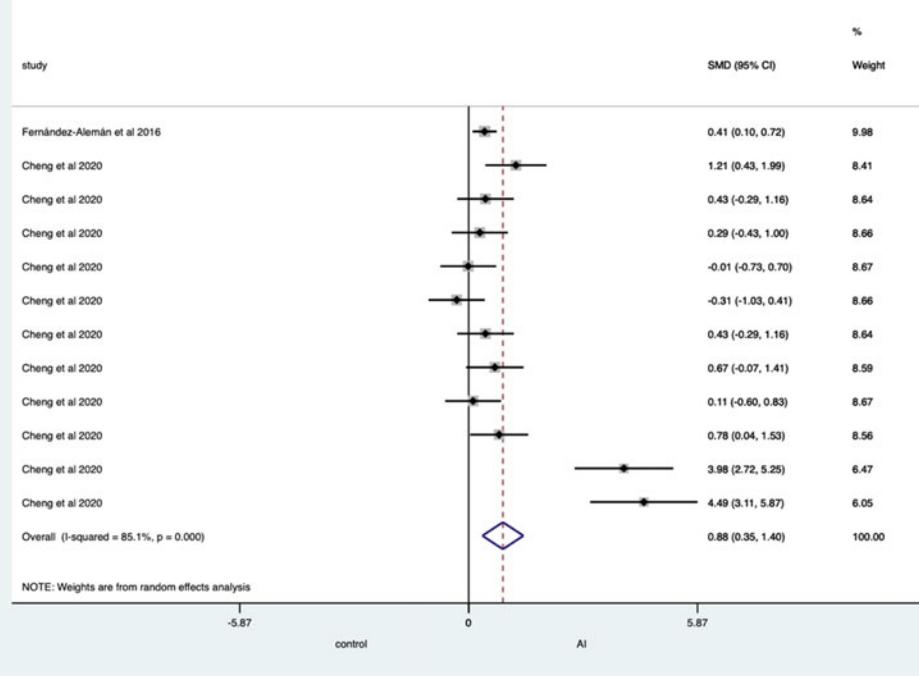
Heterogeneity chi-squared = **158.28** (d.f. = 11) p = **0.000**
 I-squared (variation in WMD attributable to heterogeneity) = **93.1%**
 Estimate of between-study variance Tau-squared = **14.0113**

Test of WMD=0 : z= **3.25** p = **0.001**

Fig. 2 The weighted mean difference using a random effects model supports a 4.789 times improvement of AI in various aspects of medical education over comparator. This would mean there appears to be a 479% improvement

in favor of AI in medical education compared with conventional methods alone suggesting promise in this area and worth exploring

SMD



SMD

Study	SMD	[95% Conf. Interval]	% Weight
Fernández-Alemán e	0.407	0.097 0.718	9.98
Cheng et al 2020	1.210	0.428 1.993	8.41
Cheng et al 2020	0.432	-0.293 1.156	8.64
Cheng et al 2020	0.285	-0.434 1.005	8.66
Cheng et al 2020	-0.012	-0.728 0.704	8.67
Cheng et al 2020	-0.308	-1.028 0.412	8.66
Cheng et al 2020	0.432	-0.293 1.156	8.64
Cheng et al 2020	0.672	-0.065 1.409	8.59
Cheng et al 2020	0.113	-0.603 0.829	8.67
Cheng et al 2020	0.782	0.037 1.526	8.56
Cheng et al 2020	3.981	2.717 5.246	6.47
Cheng et al 2020	4.492	3.115 5.869	6.05
D+L pooled SMD	0.878	0.352 1.404	100.00

Heterogeneity chi-squared = **73.75** (d.f. = 11) p = **0.000**
 I-squared (variation in SMD attributable to heterogeneity) = **85.1%**
 Estimate of between-study variance Tau-squared = **0.6964**

Test of SMD=0 : z= **3.27** p = **0.001**

Fig. 3 The standardized mean difference meta-analysis showing a significant benefit for improvement with AI applied to various aspects of medical education. Moderate

degree of heterogeneity is observed with an I^2 of 85% with an 88% improvement for using AI over conventional methods alone

ROM

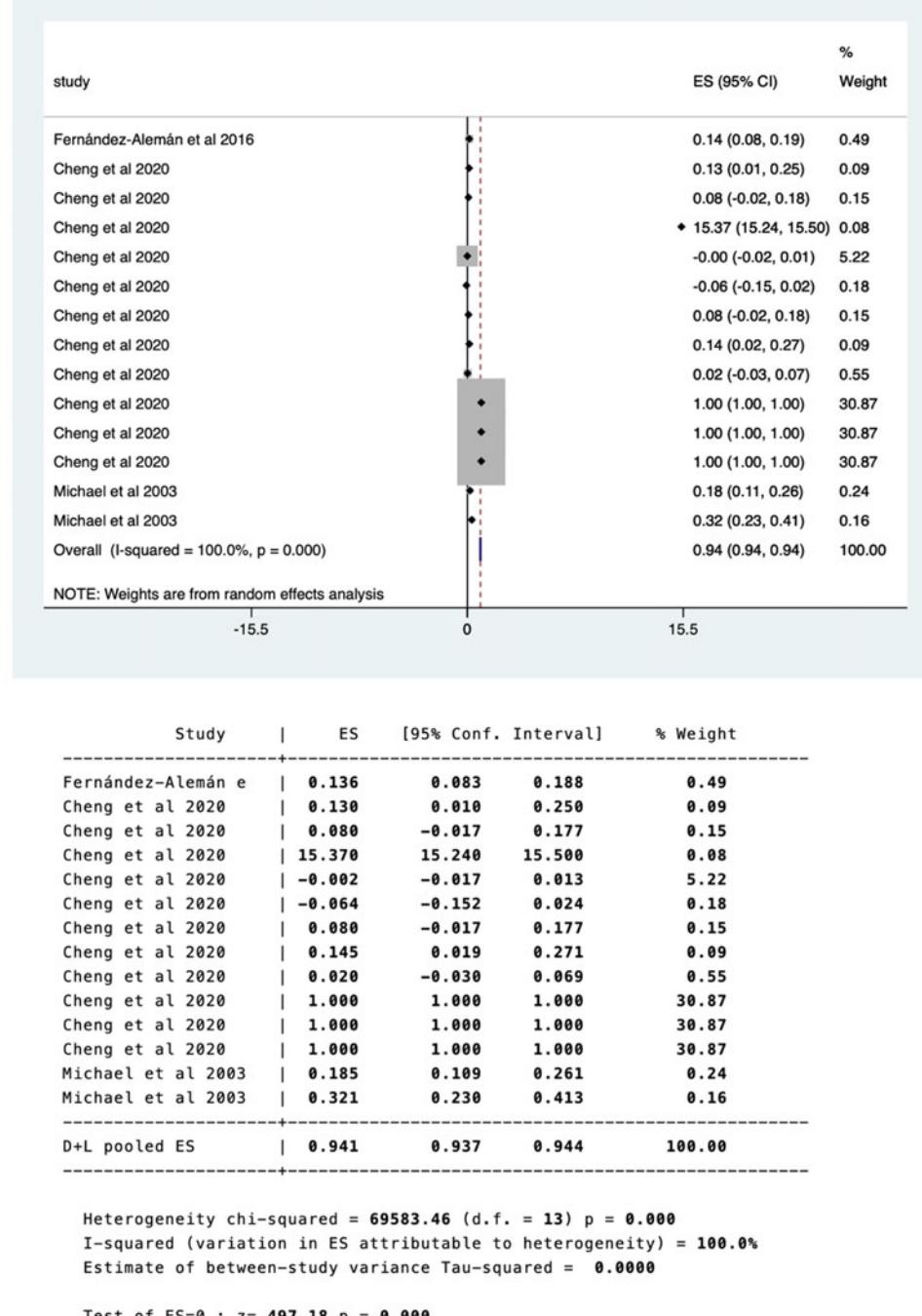


Fig. 4 Ratio of means meta-analysis suggesting significance, but with a high study heterogeneity between study I^2 , but a 94% improvement for using AI in medical education compared to conventional methods alone

sensitivity, and specificity. There is still a chasm that must be crossed between artificial intelligence model explainability and how this impacts model delivery. Skepticism still surrounds the black-box nature of these models. However, it is likely that as evidence continues to accumulate and newer more robust models/approaches to analyze their behavior are developed, the gap that exists between model selection and implementation in education will also close.

Currently, the global community remains open-minded about AI for education but continues to research opportunities to stay in touch with developments in various spaces. In our meta-analysis, study heterogeneity ranged from 85% to 100% commensurate with the model tested. We presented all models 1) for transparency; 2) to demonstrate that this approach may benefit from a standardized regimen as there is yet to be an established consensus on how to meta-analyze projects involving machine learning and artificial intelligence within the global community; and 3) to circumvent and reduce model selection bias as the ratio of means was also robust to missing values of non-reported standard deviations. Thus, selecting a single model may not be the best approach as it is also bias prone.

Proposing a New Method of Unbiased Reporting of Meta-Analysis of Improvement in AI Studies

In our study, the weighted mean difference approach demonstrated a highly significant effect size estimate using a random effects model of 479% improvement compared to not using artificial intelligence and presented the greatest effect size estimate among the three models. In contrast, the ratio of means and the standardized mean difference both showed near-comparable effect size estimates of 94.1% versus 87.8% improvement, respectively, using a random effects model.

Both SMD and WMD factor into their calculations of standard deviations and tend to eliminate absent values within a non-imputed and incomplete dataset. We propose to the community looking at improvement domains to report

all significant effect size estimates from models as either in a standardized or nonstandardized descending order ratio – call it “effect size index of improvement.” For instance, our study presents a ratio in descending order of ES as follows; $\text{WMD}_{\text{RE}}: \text{ROM}_{\text{RE}}: \text{SMD}_{\text{RE}}$ and could either be non-standardised [5.80] or standardised over the highest number of studies used by the models meta-analysed ($\text{WMD}_{\text{RE}}: \text{ROM}_{\text{RE}}: \text{SMD}_{\text{RE}})/14 = [0.414]$, here _{RE} stands for Random Effects, but could equally be the Fixed Effect _{FE} model used). This may eliminate the bias in reporting just the significant model, but also allows the research community to better assess the impact of improvement reported for that AI. Future evaluation of such a metric will help identify the degree of usefulness of a particular AI tool in a meta-analysis and eliminate reporting or model selection bias. There will also be a need to globally standardize protocols for conducting studies for AI in medical education to reduce the degree of study heterogeneity across borders as previously discussed in a recent study [90].

Nonetheless, the fact that early meta-analyses is revealing significant improvement suggests that some perceived benefit does exist for the medical education community to continue to explore the use of AI platforms over the years to come.

The State of Medical and Surgical Training and the Use of AI in Simulation

While surgical training paradigms have scarcely changed since their inception by Halsted at Johns Hopkins over a century ago, and have produced leading surgical personalities of eminence that have left their mark on the wider community [48, 49], this paradigm is now widely seen as in need of a radical overhaul. Trainees now work fewer hours than their predecessors and need to learn a wider variety of surgical techniques and keep up with advancements in surgical theory that are increasing at an ever-faster pace [50]. This has led to calls for change in surgical curricula from junior surgeons and their trainers [51–54], with a great emphasis placed on surgical skill acquisition outside the theatres [55–64]. Several studies have

shown the efficacy of simulation in enhancing the speed of skill acquisition in surgery [65, 66].

Simulation training paradigms vary. Simulation training in surgery can take myriad forms, ranging from training on anatomic replicas of the organs in question, actors, and in a more recent development, emerging novel technologies such as Virtual and Augmented Reality [48, 67–72]. While AI is finding increasing utilisation in many healthcare domains [73] as explained in this chapter, the use of AI in surgical skill acquisition is still in its infancy and there are very few wide-scale applications thereof in surgical training today. Research into AI in non-surgical and non-medical education dates back several decades [74, 75]. Examples thereof include AI-assisted tutors in computer science, mathematics, basic sciences such as physics, and various other domains [75–81]. An example of large-scale deployment of an AI-assisted tutoring program is the Pittsburgh Urban Mathematics Project (PUMP), which was used to help 470 middle school students improve their mathematics skills, and was shown to be effective in helping these students achieve higher grades than their counterparts [82]. Numerous other AI-assisted tutoring programs have shown their efficacy in real-world scenarios and find wide deployment in the classroom. Higher education institutions have recently begun to recognise the importance of educating their students in the field of AI; two examples hereof include the Massachusetts Institute of Technology, which aims to invest \$1 billion to build an AI-focused college, and the University of Oxford, which will institute its first new college in 30 years, Reuben College, with a focus on new technologies including AI [83, 84].

The market size for AI in education is estimated to have a value of about \$6 billion by 2024, with a growth rate that is projected to be 45% [85]. Supranational bodies such as the EU have also recognised the potential impact on Education. The European Commission Science for Policy recognises that “in the next years AI will change learning, teaching, and education,” and declaring it a “political priority” [86]. The increasing availability of AI-assisted education tools proving their utility in real-world classrooms [87], makes it imperative that surgical educators add

this valuable tool to their education arsenal. We need to evolve beyond the limited and few-and-far in between deployment of basic intelligent tutoring systems devised in the last decade of the past century [88], to make use of the astounding advancement in AI in order to better educate our surgical junior trainees, and look towards specialties that have begun to utilise AI in education such as radiology [89]. We hope that this chapter illustrates where the use of AI in surgical education currently stands, and will provide an impetus as to where AI deployment can improve surgical training.

AI in Medical Specialties for Education

In an era of holistic approaches to care, there is a shift toward more personalized healthcare. The ability of an AI to learn without explicit instruction has multiple applications in personalized education in medicine and specifically in the field of radiology. Much attention has been focused on teaching radiology trainees about AI with fear that radiologist roles will be usurped by an AI, but the reverse is beginning to gain traction with multiple institutions investing in using AI to aid teaching radiology trainees to augment their diagnostic capabilities. One study considered AI-augmented radiology education for precision medicine where specific needs-based instruction was tailored to individuals [89]. Currently, multiple successful AI radiological applications have ranged from (i) abnormality detection, (ii) anatomic segmentation, (iii) image quality assessment, (iv) natural language processing (NLP), (v) improvement of protocols and worklists, etc. [90]. The development of intelligent teaching files, which are case-based repositories for radiology, can also be integrated with adaptive learning platforms to facilitate personalized learning [89]. One tool that has been developed is the Adaptive Radiology Interpretation and Education System (ARIES). ARIES is a Bayesian network teaching platform tool that has been effective at distinguishing Neuro-Behcet’s disease from other diagnoses in the basal ganglia neuroradiology network. Table 1 summarizes some of these AI agents and how they have been used to implement medical education.

AI in Ophthalmology, Dermatology, Cardiology, Gastroenterology, and Rheumatology

Image and video analysis in ophthalmology and multisystem interface dermatology education have benefitted from the use of deep learning as a tool for quantitative assessment of pathology. Infectious diseases education has also leveraged intelligent tutoring systems like MYCIN and GUIDON to deliver education on diagnosis using a computer-based instruction platform. Cardiology and cardiovascular physiology have benefitted from using AI for educating learners on echocardiographic studies and electroencephalograms with page-turning architectures developed to achieve these objectives using platforms like CARDIOLOG and CIRCSIM tutor. In gastroenterology, artificial intelligence systems have been developed as e-Learning platforms using preprocessed clinical features and machine learning models such as support-vector machines, artificial neural networks, and decision trees. In general, medical case vignettes for rheumatology using AI-powered web-delivered multimedia-based training platforms have also been reported.

AI in Obstetrics and Gynecology, General Surgery, Orthopedic, and Neurosurgery Education

Obstetrics and gynecology as well as subspecialty general surgery have also benefited from these AI-augmented technologies such as laparoscopic trainers and intelligent tutoring platforms. In system-based learning, locomotor system education requiring descriptive anatomy has also benefited from artificial intelligence augmentation. AI has also been used in orthopedic surgical education and for fracture classification and didactic analysis such as the TELEOS project for screw fixation, AI system for hemilaminectomies, virtual operative assistance, and virtual reality systems for procedural teaching and performance assessments [91].

AI for Surgical Performance Assessment

Surgery is a discipline that carries inherent risk. For surgeons in training, accurate and reliable assessment of performance is therefore critical. However, while surgery has often been compared to other

disciplines with similar levels of risk, such as aviation or motor-racing, the assessment of surgeons is much less sophisticated than in these fields. Seminal work from the Michigan Bariatric Surgery Collaborative in 2013 demonstrated the importance of assessing surgical performance, finding a direct relationship between performance and patient outcomes [92]. Performance assessment should not be limited to surgeons in training only. While surgeons may complete training and are deemed capable of independent practice, the continuous introduction of novel techniques and technology demands the need for reliable and objective measures of performance. The development of reliable and efficient performance scoring systems could therefore have far-reaching consequences including the accurate determination of learning curves for novel techniques or technology, revalidation of surgeons, determining patient selection of surgeons, and selection of doctors for surgical training programs.

Surgical training is broadly based on an apprenticeship model, and as a result, performance assessment is still largely delivered by the expert mentor. However, means do exist for more objective measures of performance assessment. Observer rating tools such as the Objective Structured Assessment of Technical Skills (OSATS) [93, 94] require experts to rate performance of tasks on a procedure-based checklist as well as a global rating scale. Such techniques have also been adapted for minimally invasive techniques [95, 96] or to be procedure-specific [97]. However, these tools require expert assessment, are time-consuming, and are prone to rater bias.

The modern-day operating room now offers a wealth of data including sensor, kinematic, and video data. AI allows surgeons to capitalize on these heterogeneous datasets and move away from blunt indicators of performance such as operative time and postoperative outcomes. The use of AI has the promise of rapid, automated, objective, and reproducible surgical performance assessment which would allow continuous feedback limited only by hardware and without the need of an expert assessor.

Current surgical robotic platforms allow precise mapping of the movement of its instrument arms, and tools such as the dVLogger (Intuitive Surgical,

Sunnyvale, California) have allowed the recording of kinematics and video data. The use of these kinematic datasets can be leveraged to provide objective and constructive feedback to the surgeon. The simplest way to translate this kinematic data is its conversion into global movement features. These range from time to completion, speed and number of hand movements, path length, force, and torque. These metrics have been successfully processed via a variety of machine learning techniques such as k-nearest neighbors, logistic regression, and support-vector machine to achieve good accuracy in determining expertise level in robotic benchtop tasks [98]. While they are able to give a generic measure of technical ability, the clinical utility to a training surgeon is limited. It is more important for surgical trainees to understand why they have been classified as a novice than simply that they are one.

Deep learning allows models consisting of many processing layers to process and learn from the input data achieving state-of-the-art results in pattern recognition [99]. As a result, deep learning is an ideal candidate for feature extraction. Use of convolutional neural networks (CNN) can translate kinematic data into interpretable feedback for surgeons. The utilization of a class activation map can inform which part of the task plays a significant part of the model's decision when evaluating the skill level of the surgeon [100]. Therefore, surgeons are able to identify which behaviors are specific to a certain skill level and thus identify good or bad behaviors. Moreover, CNNs have also successfully predicted OSATS scores, and class activation maps have been similarly applied in order to identify key aspects of the task influencing the score [100].

Although assessment of surgical performance using kinematics has largely been in benchtop tasks, they have also been applied to real-life surgery. A study investigating automated performance metrics (APM), motion-tracking and events metrics, in 100 robot-assisted radical prostatectomies (RARP) has shown experts perform the four key steps of the procedure faster with less idle time and travelling less distance [101]. Further clinical significance of measuring APMs is in its applicability to patient outcomes. The use of APMs alongside machine learning algorithms

has been shown to be able to predict length of stay [101] and also long-term outcomes such as continence following RARP [102].

Data used for surgical performance assessment is not restricted to kinematics only. Surgical video datasets of laparoscopic cholecystectomies have been processed to allow spatial detection of surgical tools based on the region-based convolutional neural network, Faster R-CNN [103]. This output allowed qualitative and quantitative analysis of tool movements such as usage patterns and motion economy in order to predict surgical skill and performance that was consistent with expert ratings.

The application of artificial intelligence for surgical performance assessment shows significant promise. Surgeons of the future will be able to obtain automatic objective analysis of intraoperative performance, which will accelerate surgical education and allow surgeons to advance up learning curves, first as a trainee and then continually throughout their career.

The use of AI for surgical performance assessment, however, is still in its relative infancy. A large proportion of work has been undertaken on the JIGSAWS dataset [104], an annotated open-source dataset of kinematic data, video data, and manual annotations of eight surgeons performing various benchtop tasks. Future advances in AI-driven surgical performance assessment are dependent on the increasing availability of such datasets. The advent of surgical video datasets such as Cholec-80 and Lapsig-300 is promising and may provide the impetus for more units to develop their own datasets [105, 106]. The focus for future AI assessment tools, now, must not only be accuracy but also provision of targeted and constructive feedback for surgical trainees, which is clinically relevant.

AI for Precision Examination and Ethico-Legal Aspects of Medical Education and Autonomous Robotic Surgical Regulation

Another area is precision examination and exam practice, where artificial intelligence is being designed and implemented both to mimic human actors in understanding decision-making and to help to tailor questions for revision aids. The

United States Medical Licensing Examination is one such area that teams like Beam and Colleagues at the Department of Biomedical Informatics in the Blavatnik Institute at Harvard Medical School are developing AI models for trialing [107]. Other algorithms such as natural language processing have been considered to aid a scoring system for USMLE [108, 109]. Companies such as CloudMex AI's deep learning model were able to outperform doctors taking the USMLE exam by 10% improvement. Doctors working with CloudMex had a significant improvement in a simulated USMLE exam scores compared with average doctors taking the same exam. Group 1 were human doctors averaging 75% (range 68–81%) which was less than the CloudMedx AI score of 85%. Impressively, Group 3 human doctors along with CloudMedx AI scored 91% [110]. Undoubtedly, the combined medical intelligence of an AI with the human doctor can be beneficial to the patient and improve outcomes.

For completeness, we also briefly mention the need for legal, regulatory, and ethical frameworks for AI tools developed both for medical education and for standard development of protocols for autonomous robotic surgery. This is an area that is necessary to help identify accountability, culpability, and liability. If an autonomous agent is taught/trained to perform an independent robotic procedure and does it erroneously leading to harm, the question of who is responsible needs to be considered and a subject of considerable debate. So too is the question of who is responsible if an autonomous agent teaches a learner erroneously that then impacts negatively on medical practice, thus leading to a clinical error. Safety critical systems for AI and cybersecurity are also worth mentioning as the community becomes dependent on the digital infrastructure that an AI is built on where vulnerabilities could lead to autonomous agents being hijacked by attackers with devastating safety and economical consequences [111].

Conclusion

AI continues to be used in various medical subspecialties to augment education and has a growing impact across the board. Medically, various

subspecialties are seeing the need for artificial intelligence. This chapter's systematic review and meta-analysis show that there is a potential for AI to augment educational needs in medicine and support workforce training and performance needs. AI-powered surgical assessment also offers the potential to revolutionize surgical education by providing surgeons with continuous and reproducible feedback on their intraoperative performance. There have been several discussions about the challenges faced and the black-box nature of how some artificial intelligence models like neural networks work in their classification. For strong opponents of AI based on explainability for most applications including medical education, there is a counterargument that just because we do not completely understand how our brains work perhaps does not mean we should not use our brains to solve problems. To which the opposing view would be that just because we know how to and can use our brains but do not completely understand how does not mean we should stop trying to understand how it works. Both points of view are valid, and in the context of education where we seek knowledge by learning, understanding, and implementing what has been learned, we must appreciate how both sides of the coin work by staying open-minded. As we use machines to enhance education and its delivery, we must also start to learn from how these machines are behaving to help design better systems to benefit society. The interaction must be bidirectional allowing us to improve ourselves from an educational perspective while improving and discovering better AI to further augment education.

References

1. Fulton JF. History of medical education. Br Med J. 1953;2:457.
2. Heffernan GE. Discours de La Methode/Discourse on the method: a bilingual edition with an interpretive essay. Notre Dame: University of Notre Dame Press; 1994.
3. Vivarès F. Automata 1742. <https://pictures.royalsociety.org/image-rs-11869>
4. Wood G. Living dolls: a magical history of the quest for mechanical life. The Guardian. 2002.

5. Stock J, Esposito M, Lanteri V. Urologic robotic surgery – current clinical urology. Humana Press; 2008.
6. Fryer D, Marshall J. The motives of Jacques de Vaucanson. *Technol Cult*. 1979;20(2):257–69.
7. Ashrafiyan H, Darzi A, Athanasiou T. A novel modification of the Turing test for artificial intelligence and robotics in healthcare. *Int J Med Robot*. 2015;11(1):38–43.
8. Callaway E. ‘It will change everything’: DeepMind’s AI makes gigantic leap in solving protein structures. *Nature*. 2020;588:203–4.
9. Silver D, Huang A, Maddison C, et al. Mastering the game of go with deep neural networks and tree search. *Nature*. 2016;529:484–9.
10. Hunt E. Tay, Microsoft’s AI chatbot, gets a crash course in racism from Twitter. *The Guardian*. 2016.
11. Schröder H, Henke A, Stieger L, Beckers S, Biermann H, Rossaint R, Sopka S. Influence of learning styles on the practical performance after the four-step basic life support training approach – an observational cohort study. *PLoS One*. 2017;12(5):e0178210.
12. Bergsteiner H, Avery G, Neumann R. Kolb’s experiential learning model: critique from a modelling perspective. *Stud Contin Educ*. 2010;32(1):29–46.
13. Kolb D. Experiential learning: experience as the source of learning and development. Pearson FT Press PTG; 2015.
14. Honey P, Mumford A. The learning styles helper’s guide. Maidenhead: Peter Honey; 2000.
15. Honey P, Mumford A. The manual of learning styles. Maidenhead: Peter Honey; 1986.
16. Gregorc A, Butler K. Learning is a matter of style. *Vocat Educ J*. 1984;59(3):27–9.
17. Baig M, Ahmad M. Learning with a style: the role of learning styles and models in academic success. *Eur Acad Res*. 2016;4(8):6695–705.
18. Bajaj R, Sharma V. Smart education with artificial intelligence based determination of learning styles. *Procedia Comput Sci*. 2018;132:834–42.
19. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan – a web and mobile app for systematic reviews. *Syst Rev*. 2016;5:210.
20. Bastardot F, Gachoud D. Visual diagnosis: between medical education and advances in artificial intelligence. *Rev Med Suisse*. 2019;15(672):2145–9.
21. van der Niet A, Bleakley A. Where medical education meets artificial intelligence: ‘does technology care?’? *Med Educ*. 2020;55(1):30–6.
22. Lindqvister A, et al. AI-RADS: an artificial intelligence curriculum for residents. *Acad Radiol*. 2020; S1076-6332(20)30556-0.
23. Carin L. On artificial intelligence and deep learning within medical education. *Acad Med*. 2020;95(11S):S1076-6332(20)30556-0.
24. Clancey W. GUIDON. *J Comput-Based Instruct*. 1983;10(1 & 2):8–15.
25. Bourlas P, Giakoumakis E, Koutsouris D, Papakonstantinou G, Tsanakas P. The CARDIO-LOGOS system for ECG training and diagnosis. *Technol Health Care*. 1996;3(4):279–85.
26. Voss G, et al. LAHYSTOTRAIN intelligent training system for laparoscopy and hysteroscopy. *Stud Health Technol Inform*. 2000;70:359–64.
27. Stasiu RK, et al. Teaching of electrocardiogram interpretation guided by a tutorial expert. In: Proceedings 14th IEEE symposium on computer-based medical systems; 14th IEEE symposium on computer-based medical systems. 2001. p. 487–92.
28. Kintsch W. The potential of latent semantic analysis for machine grading of clinical case summaries. *J Biomed Inform*. 2002;35(1):3–7.
29. Caudell T, et al. Virtual patient simulator for distributed collaborative medical education. *Anat Rec B New Anat*. 270(1):23–9.
30. Crowley R, Medvedeva O. A general architecture for intelligent tutoring of diagnostic classification problem solving. In: AMIA annual symposium proceedings. AMIA Symposium; 2003. p. 185–9.
31. Michael J, Rovick A, Glass M, Zhou Y, Evens M. Learning from a computer tutor with natural language capabilities. *Interact Learn Environ*. 2003;11(3):233–62.
32. McFadden P, Crim A. Comparison of the effectiveness of interactive didactic lecture versus online simulation-based CME programs directed at improving the diagnostic capabilities of primary care practitioners. *J Contin Educ Heal Prof*. 2016;36(1):32–7.
33. Khumrin P, Ryan A, Judd T, Verspoor K. Diagnostic machine learning models for acute abdominal pain: towards an e-learning tool for medical students. *Stud Health Technol Inform*. 2017;245:447–51.
34. Chen H, Gangaram V, Shih G. Developing a more responsive radiology resident dashboard. *J Digit Imaging*. 2019;32(1):81–90.
35. Cheng C, Chen CC, Fu CY, et al. Artificial intelligence-based education assists medical students’ interpretation of hip fracture. *Insights Imaging*. 2020;11:119.
36. El Saadawi GM, Tseytlin E, Legowski E, et al. A natural language intelligent tutoring system for training pathologists: implementation and evaluation. *Adv Health Sci Educ*. 2008;13:709–22.
37. Chieu VM, Luengo V, Vadcard L, Tonetti J. Student modeling in orthopedic surgery training: exploiting symbiosis between temporal Bayesian networks and fine-grained didactic analysis. *Int J Artif Intell Educ*. 2010;20:269–301.
38. Fernández-Aleman JLL-GL, González-Sequeros O, Jayne C, López-Jiménez JJ, Toval A. The evaluation of i-SIDRA – a tool for intelligent feedback – in a course on the anatomy of the locomotor system. *Int J Med Inform*. 2016;94:172–81.
39. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. *JMIR Med Educ*. 2019;5(2):e16048.
40. Lang J, Repp H. Artificial intelligence in medical education and the meaning of interaction with natural intelligence – an interdisciplinary approach. *GMS J Med Educ*. 2020;37(6):Doc59.

41. Frize M, Frasson C. Decision-support and intelligent tutoring systems in medical education. *Clin Invest Med.* 2000;23(4):266–9.
42. Masters K. Artificial intelligence in medical education. *Med Teach.* 2019;41(9):976–80.
43. Holden MS, Xia S, Lia H, et al. Machine learning methods for automated technical skills assessment with instructional feedback in ultrasound-guided interventions. *Int J CARS.* 2019;14:1993–2003.
44. Mirchi N, Bissonnette V, Yilmaz R, Ledwos N, Winkler-Schwartz A, Del Maestro RF. The virtual operative assistant: an explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PLoS One.* 2020;15(2):e0229596.
45. Bissonnette V, Mirchi N, Ledwos N, Alsidieri G, Winkler-Schwartz A, Del Maestro R. Artificial intelligence distinguishes surgical training levels in a virtual reality spinal task 2019 2019-12-4. e127 p.
46. Winkler-Schwartz ABV, Mirchi N, Ponnudurai N, Yilmaz R, Ledwos N, Siyar S, Azarnoush H, Karlik B, Del Maestro RF. Artificial intelligence in medical education: best practices using machine learning to assess surgical expertise in virtual reality simulation. *J Surg Educ.* 2019;76:6.
47. Wang Z, Majewicz F. Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. *Int J Comput Assist Radiol Surg.* 2018;13(12):1959–70.
48. Haluck R, et al. Computers and virtual reality for surgical education in the 21st century. *Arch Surg.* 2000;135(7):786–92.
49. Rombeau JL, Goldberg A, Loveland-Jones C. Surgical mentoring: building tomorrow's leaders. Heidelberg: Springer; 2010.
50. Lyon P. A model of teaching and learning in the operating theatre. *Med Educ.* 2004;38(12):1278–87.
51. Haase J, et al. Neurosurgical training: more hours needed or a new learning culture? *Surg Neurol.* 2009;72(1):89–95.
52. Regelsberger J, et al. Training microneurosurgery -four years experiences with an *in vivo* model. *Cent Eur Neurosurg.* 2011;72(4):192–5.
53. Rodriguez-Paz J, et al. Beyond “see one, do one, teach one”: toward a different training paradigm. *Qual Saf Health Care.* 2009;18(1):63–8.
54. Sooriakumaran P. Is UK surgical training in crisis? A trainee's perspective. *Int J Surg.* 2004;2(3):127.
55. Burkhardt J, et al. Neurosurgical education in Europe and the United States of America. *Neurosurg Rev.* 2010;33(4):409–17.
56. Morgan MK, et al. The neurosurgical training curriculum in Australia and New Zealand is changing. Why? *J Clin Neurosci.* 2005;12(2):115–8.
57. Sure U, Miller D, Bozinov O. Neurosurgical training in Europe, problems and possible solutions. *Surg Neurol.* 2007;67(6):626–8.
58. Aggarwal R, Darzi A. Competency-based training and practice—what does it really mean? *J Am Coll Surg.* 2007;205(1):192–3.
59. Morris C. Facilitating learning in the workplace. *Br J Hosp Med (Lond).* 2010;71(1):48–50.
60. Hamamcioglu MK, et al. A laboratory training model in fresh cadaveric sheep brain for microneurosurgical dissection of cranial nerves in posterior fossa. *Br J Neurosurg.* 2008;22(6):769–71.
61. Regelsberger J, et al. In vivo porcine training model for cranial neurosurgery. *Neurosurg Rev.* 2015;38(1):157–63.
62. Salma A, Chow A, Ammirati M. Setting up a micro-neurosurgical skull base lab: technical and operational considerations. *Neurosurg Rev.* 2011;34(3):317–26.
63. Takeuchi M, et al. A new training method to improve deep microsurgical skills using a mannequin head. *Microsurgery.* 2008;28(3):168–70.
64. Turan Suslu H, Tatarli N, Hicdonmez T, Borekci A. A laboratory training model using fresh sheep spines for pedicular screw fixation. *Br J Neurosurg.* 2012;26(2):252–4.
65. Amr A, et al. Testing the efficacy of simulation in neurosurgical education: first results of the SENSE trial. *Neurosurgery.* 2017;64:223–4.
66. Ganju A, et al. The role of simulation in neurosurgical education: a survey of 99 United States neurosurgery program directors. *World Neurosurg.* 2013;80(5):1–8.
67. Berhouma M, et al. Shortening the learning curve in endoscopic endonasal skull base surgery: a reproducible polymer tumor model for the trans-sphenoidal trans-tubercular approach to retro-infundibular tumors. *Clin Neurol Neurosurg.* 2013;115(9):1635–41.
68. Menovsky T. A human skull cast model for training of intracranial microneurosurgical skills. *Microsurgery.* 2000;20(7):311–3.
69. Choudhury N, et al. Fundamentals of neurosurgery: virtual reality tasks for training and evaluation of technical skills. *World Neurosurg.* 2013;80(5):9–19.
70. Paloc C, et al. Virtual reality surgical training and assessment system. *Int Congr Ser.* 2001;1230:210–7.
71. Robison RA, Liu CY, Apuzzo ML. Man, mind, and machine: the past and future of virtual reality simulation in neurologic surgery. *World Neurosurg.* 2011;76(5):419–30.
72. Schmitt PJ, Agarwal N, Prestigiacomo CJ. From planes to brains: parallels between military development of virtual reality environments and virtual neurological surgery. *World Neurosurg.* 2012;78(3–4):214–9.
73. Jiang F, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol.* 2017;2(4):230–43.
74. Beck J, et al. Applications of AI in education. *Crossroads.* 1996;3(1):11–5.
75. Katz A, Ochs J. Profiling student activities with the Smithtown discovery world program. *Soc Sci Comput Rev.* 1993;11(3):366–78.

76. El Agha M, et al. SQL tutor for novice students. *Int J Acad Inf Syst Res.* 2018;2(2):1–7.
77. Qwaider S, Abu-Naser SS. Excel intelligent tutoring system. *Int J Acad Inf Syst Res.* 2018;2(2):8–18.
78. VanLehn K, Lynch C, Schultz K, Shapiro JA, Shelby RH, Taylor L, et al. The Andes physics tutoring system: lessons learned. *Int J Artif Intell Educ.* 2005;15(3):147–204.
79. VanLehn K, van de Sande B, Shelby R, Gershman S. The Andes physics tutoring system: an experiment in freedom. In: Nkambou R, Bourdeau J, Mizoguchi R, editors. *Advances in intelligent tutoring systems studies in computational intelligence.* Berlin/Heidelberg: Springer; 2010. p. 308.
80. Luckin R, et al. *Intelligence unleashed: an argument for AI in education.* London: Pearson; 2016.
81. Porayska-Pomsta K. AI in Education as a methodology for enabling educational evidence-based practice. Workshop on Les Contes du Mariage: should AI stay married to Ed? 2015. p. 52–61.
82. Koedinger KR, et al. Intelligent tutoring goes to school in the big city. *Int J Artif Intell Educ (IJAIED).* 1997;8:30–43.
83. Office MN. MIT reshapes itself to shape the future 2018. <http://news.mit.edu/2018/mit-reshapes-itself-stephen-schwarzman-college-of-computing-1015>
84. Office ON. Oxford unveils plans for new graduate college 2018. <http://www.ox.ac.uk/news/2018-12-07-oxford-unveils-plans-new-graduate-college>
85. Bhutani A, Wadhwan P. Artificial Intelligence (AI) in education market size 2018. <https://www.gminsights.com/industry-analysis/artificial-intelligence-ai-in-education-market>
86. Tuomi I. JRC science for policy report: the impact of artificial intelligence on learning, teaching, and education. Joint Research Centre (European Commission). 2018.
87. Mitrovic A, et al. ASPIRE: an authoring system and deployment environment for constraint-based tutors. *Int J Artif Intell Educ (IJAIED).* 2009;19(2):155–88.
88. Evens M, et al. CIRCSIM-tutor: an intelligent tutoring system using natural language dialogue. In: Proceedings of the fifth conference on applied natural language processing: descriptions of system demonstrations and videos: Association for Computational Linguistics. p. 13–4.
89. Duong M, et al. Artificial intelligence for precision education in radiology. *Br J Radiol.* 2019;92(1103):20190389.
90. Lakhani P, et al. Machine learning in radiology: applications beyond image interpretation. *J Am Coll Radiol.* 2018;15:350–9.
91. Davids J, Manivannan S, Darzi A, et al. Simulation for skills training in neurosurgery: a systematic review, meta-analysis, and analysis of progressive scholarly acceptance. *Neurosurg Rev.* 2020.
92. Birkmeyer J, et al. Surgical skill and complication rates after bariatric surgery. *N Engl J Med.* 2013;369(15):1434–42.
93. Martin J, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg.* 1997;84:273–8.
94. Reznick R, et al. Testing technical skill via an innovative “bench station” examination. *Am J Surg.* 1997;173:226–30.
95. Goh A, et al. Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. *J Urol.* 2012;187(1):247–52.
96. Vassiliou M, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg.* 2005;190:107–13.
97. Insel A, et al. The development of an objective model to assess arthroscopic performance. *J Bone Joint Surg Am.* 2009;91(9):2287–95.
98. Fard M, et al. Machine learning approach for skill evaluation in robotic-assisted surgery. In: *Proceedings of the world congress on engineering and computer science.* 2016.
99. LeCun Y, et al. Deep learning. *Nature.* 2015;521:436–44.
100. Fawaz H, et al. Evaluating surgical skills from kinematic data using convolutional neural networks. In: *MICCAI 2018: Medical image computing and computer assisted intervention,* vol. 11073. 2018. p. 214–21.
101. Hung A, et al. Utilizing machine learning and automated performance metrics to evaluate robot-assisted radical prostatectomy performance and predict outcomes. *J Endourol.* 2018;32(5):438–44.
102. Hung A, et al. A deep-learning model using automated performance metrics and clinical features to predict urinary continence recovery after robot-assisted radical prostatectomy. *BJU Int.* 2019;124(3):487–95.
103. Jin A, et al. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV).* 2018. p. 691–9.
104. Gao Y, et al. JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS): a surgical activity dataset for human motion modeling. 2014.
105. Twinanda A, et al. EndoNet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE Trans Med Imaging.* 2017;36:86–97.
106. Kitaguchi D, et al. Automated laparoscopic colorectal surgery workflow recognition using artificial intelligence: experimental research. *Int J Surg.* 2020;79:88–94.
107. Pesheva E, Menting A. HMS communications; Science and Technology; [Internet]. In: Gazette TH, editor. *The Harvard Gazette.* 2019. [cited 2021]. <https://news.harvard.edu/gazette/story/2019/04/at-harvard-adding-ai-to-m-d/>
108. Salt J, Harik P, Barone MA. Leveraging natural language processing: toward computer-assisted

- scoring of patient notes in the USMLE step 2 clinical skills exam. *Acad Med.* 2019;94(3):314–6.
109. Wartman SA, Combs CD. Reimagining medical education in the age of AI. *AMA J Ethics.* 2019;21(2):E146–52.
110. Desk AN. CloudMedx Clinical AI outperforms human doctors on a US medical exam: AI Authority; 2019. <https://aithority.com/machine-learning/neural-networks/deep-learning/cloudmedx-clinical-ai-outperforms-human-doctors-on-a-us-medical-exam/>
111. O’Sullivan S, Nevejans N, Allen C, et al. Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. *Int J Med Robotics Comput Assist Surg.* 2019;15:e1968.



AIM and Evolutionary Theory

23

Jonathan R. Goodman and Nicolai Wohns

Contents

Introduction	342
Mathematical Oncology	343
Prediction of Cancer Risk	344
Precision Medicine	344
Future Directions	345
Infectious Disease	345
Drug Resistance	346
Drug Design	347
Emerging Pathogens	347
Avoiding the Desynthesis	348
References	348

Abstract

Artificial intelligence (AI) applied to the genome sciences has the potential to revolutionize healthcare. Yet to fully harness the predictive power of AI, at least in the fields of oncology and infectious disease, evolutionary

theory must be brought to bear. In oncology, AI is uniquely suited to analyze the complex lattice-work of correlations among the many genomic and environmental influences that constitute cancer risk. It also makes possible the evolutionary theory-inspired concepts of next-generation cancer treatment, such as evolutionary traps, adaptive therapy, and treatment vaccination. In infectious disease, AI promises the rapid diagnosis of a pathogen's current drug resistance profile, as well as the prediction of its potential to develop resistance. Using anticipatory diagnostics, drug regimens can be tailored to probabilistically channel pathogens toward less resistance-prone genotypes to avoid the emergence of resistance. Advanced computational methods are also used in antimicrobial drug design and to anticipate

J. R. Goodman (✉)
Leverhulme Centre for Human Evolutionary Studies,
University of Cambridge, Cambridge, UK

Darwin College, University of Cambridge, Cambridge,
UK
e-mail: jrg74@cam.ac.uk

N. Wohns
Department of Philosophy, University of Washington,
Seattle, WA, USA
e-mail: nwohns@uw.edu

outbreaks of infectious disease and the evolution of epidemics, such as the SARS-CoV-2 pandemic. In detailing these advances, we discuss illustrative examples of the productive collaboration of data scientists, evolutionary theorists, epidemiologists, and clinicians. In addition, we briefly note the dangers of over-reliance on advanced computational tools that involve “black box” algorithms and question whether they undermine the synthesis of Mendelian genetics and Darwinian theory. Yet, insofar as evolutionary theory is used for hypothesis algorithm development and AI for data creation and analysis, this problem may be avoided, and the potential of both will be realized.

Keywords

Evolution · Artificial intelligence · AI · Data science · Cancer · Infectious disease · COVID-19 · Mathematical oncology

Introduction

Research during the COVID-19 pandemic highlights the technological progress driving improved understanding of the evolution and etiology of infectious disease. Numerous ongoing studies and collaborations continue to track changes in the SARS-CoV-2 genome and to evaluate the frequency of new variants as they spread in human populations. Notably, the discovery of the B.1.1.7 variant in late 2020, and the consequent association with higher transmissibility, led directly to new lockdown measures in the United Kingdom and to other public health policy changes in countries across the world.

The success of these projects depends entirely upon improvements in technological sophistication: even a decade ago, this level of variant tracking would not have been possible. Data are now rapidly analyzed on novel platforms that rely, in many cases, on algorithms developed using artificial intelligence. For the first time in history, we are able to use these enormous technological capabilities to model, monitor, and potentially

anticipate how disease changes in its new ecological environment.

Yet a problem may be emerging that has been seen in other areas of evolutionary science: the focus on genomic change, measured using tools generated by artificial intelligence, is not necessarily Darwinian. Rather than making evolutionarily informed predictions about phenotypic change, many, if not all, of the ongoing studies in evolutionary epidemiology rely entirely on observing genomic change.

Understanding the implications of this problem requires a cursory overview of some of the major changes in the evolutionary sciences since Darwin first published his *Origins*. Prior to the modern synthesis, Mendelian genetics and Darwinian theory were entirely separate; they were, however, essentially united in the first half of the twentieth century, largely because of the mathematical contributions of R.A. Fisher and colleagues.

The synthesis of Mendelian genetics and Darwinian theory laid the groundwork for population genetics and predictive modelling across the biological world, leading, in the past century or so, to an explosion in output in the evolutionary sciences. Over the past 40 years, furthermore, evolutionary theorists have started applying these tools to behavioral trends, genomic change, and even cultural phenomena. Evolutionary theory provides the foundation for making predictions about how phenotypic change will occur; genomics helps us to evaluate, and to better understand, these changes.

The next-generation computational tools that now enable us to track genomic changes in a pandemic also have the potential to help us to predict phenotypic change, given our understanding of the Darwinian forces operating both on us and the pathogens we share with one another. The potential for improvement in the medical world – from public health recommendations to individual care optimization – is boundless, and for the first time, it seems that technology is outpacing our ability to use it.

A consequence of these rapid technological advancements may, however, be overreliance on outputs from algorithms used in the medical sciences. As much of the machine learning is done

“under the hood,” we depend on the data generated so much that we can only react to them, rather than predict them. Artificial intelligence needs an explicit evolutionary framework in order to capitalize on the power of predictive analytics; observing genomic change is not enough. COVID-19 is the perfect illustration of this issue: the new variants now raising alarm are doing so because they appear to be more transmissible than those with which they are competing. Yet from a Darwinian point of view, more transmissible variants are a natural product of circulation within human populations: the virus is fitter insofar as it spreads more effectively. Vaccine escape is also expected when vaccines are employed. Crowding millions of living animals together in market conditions gives viruses ample opportunity to adapt to new hosts, which increases the risk of zoonotic jumps.

And yet, on review of the vast literature developed between late 2019 and early 2021, very little focuses on Darwinian predictions, but rather on genomic observations. Researchers react with surprise to variants like B.1.1.7, which should, rather, have been uniformly expected. The virus’s ecology – namely, human populations – drives its evolutionary change, and the genomics should aid in our tracking these phenotypic changes. Instead, we find new variants that appear fitter and scramble to find out why.

It is difficult not to view this reactivity as a symptom of a “desynthesis” of Darwinian theory and genetics, in a way unraveling the efforts of the modern synthesis of a century ago. Yet, as we aim to show in this chapter, there is a great deal of research in the medical sciences that relies on the tenets of evolutionary selection – with the critical understanding that humans can and do drive the evolution of infectious and non-communicable diseases. What follows is an overview of this work – conceptual, modelling, and empirical – in the studies of cancer and infectious disease.

Mathematical Oncology

Evolutionary selection has a direct impact on the effectiveness of cancer therapy. Dhawan et al. note, for example, that in modern lung cancer

treatment, “...after failure of first line ALK-TKI [tyrosine kinase inhibitor] therapy, another ALK-TKI is administered, though collateral sensitivity is not considered” [1]. The authors challenged this practice by inducing resistance in non-small cell lung cancer cells with an *ALK* rearrangement. Cross-resistance was found for each of the four ALK-TKIs tested, and the authors noted that, despite these findings, oncologists have yet to consider cross-resistance before administering a second ALK-TKI after resistance to a first evolves.

Intuition suggests that if a therapy is likely to fail because of cross-resistance, it should be standard practice to attempt a second-line therapy that works on a different pathway or by a different mechanism than any treatment in the first line. This is not, however, yet common practice, and two factors – namely, underappreciation of the importance of selective pressures created by anti-cancer therapies and, relatedly, a lack of use of artificial intelligence to determine issues, created by evolutionary selection, such as collateral sensitivity – have, thus far, hindered an integrated approach to personalized medicine.

The purpose of this section is to highlight the importance of the distinction between natural, artificial, and unconscious selection in the evolution of drug resistance to cancer treatments and suggest that cross-disciplinary integration of evolutionary theory and data science is essential for future breakthroughs in personalized oncology [2]. To treat cancer drug resistance as a natural phenomenon is to imply it is a result of evolutionary selection, understood as evolution through differential reproduction. Yet once oncologists and researchers are made aware of the influence of their actions on the evolution of drug resistance, the process shifts from unconscious to unintentional anthropogenic selection, a subset of artificial selection. While this is becoming the dominant view in particular research groups, it is not widely appreciated, and awareness of this paradigm may drive its uptake [3].

If this is true, oncologists and researchers may recognize that each aspect of cancer treatment, including type and order of intervention, dosage, and timing, is of evolutionary and therefore

practical significance in oncology – each of which can be adjusted in line with evolutionarily informed predictions generated by artificial intelligence. Work in both modelling and the clinical setting are revealing the enormous potential of this field, sometimes called mathematical oncology [4, 5], which, as with other areas of medicine, is being revolutionized by improvements in evolutionary understanding and technological sophistication. We discuss the evidence thus far in prediction of cancer risk and precision medicine and note future avenues of interest.

Prediction of Cancer Risk

Diagnosing cancer at as an early a stage as possible has significant effects on patient prognosis [6]. A decade ago, Li et al. noted that traditionally used cancer risk models, which rely on a distinction between people without cancer, people with asymptomatic cancer, and people with symptomatic cancer, rely solely on morphology, which, in view of improvements in other areas relevant to oncology, is no longer sufficient for diagnostic purposes [7]. This, according to Li et al., makes distinguishing between people with cancer and people without cancer an important and challenging task [7].

Technological developments and improvements in understanding of genomics have enabled researchers to pinpoint particular heritable genomic mutations linked with cancer risk, including *BRCA1* and *BRCA2* in breast and ovarian cancer [8], *RB1* in retinoblastoma [9], *TP53* in multiple cancer subtypes [10], and *BCR-ABL* in chronic myeloid leukemia [11], though these, and other known heritable syndromes, are not linked with a wide range of cancers, leaving the majority of cancer subtypes without direct genomic links.

The issue of diagnosis is further complicated by the relatively recent realization that many cancers are probably a product of many genomic and environmental factors, which include the still poorly understood human microbiome, lifestyle, and persistent infection. Some viruses, notably human papillomavirus [12], have direct links with cancer, while others, such as cytomegalovirus [13], have

established but indeterminate links with disease. Even strongly correlated lifestyle variables like smoking do not have perfect correlation with any type of cancer, including lung cancer. Zur Hausen [14] notes, for example, that while smoking is common in India, lung cancer is not; lack of consumption of beef may, instead, explain this – and the identification of widespread bovine milk and meat factor infection in Western countries may be playing an underappreciated role.

Together, aside from some standout cases, the majority of cancers cannot be explained by individual factors, whether genomic, lifestyle, viral, or bacterial. Li et al. suggest, instead, that a dynamic system's approach, which accounts for these factors together, as well as tumor clonal evolution, be developed [7]. Improvements in artificial intelligence are, unlike a decade ago, making this possible: the apparently infinite number of connections among the factors described above can now be evaluated in predictive models, which will, over the coming years, allow researchers to stratify patients by a more realistic set of risks of cancer.

Precision Medicine

In *The Origin of Species*, Darwin distinguishes between artificial and natural selection by describing how a farmer selects for qualities in livestock. Animals that display the qualities and traits desirable to the farmer are allowed to reproduce; the others are not. The result, generations later, may be a group of animals with the characteristics the farmer intended. While all the biological changes are still consequences of differential reproduction, the farmer was the immediate agent of selection.

Darwin later introduced the concept of unconscious selection, distinguishing it from “methodical” (or artificial/anthropogenic) selection and from natural selection. Heiser argued that features distinguishing domesticated plants from their undomesticated ancestors may result from this unconscious process: artificial selection may produce some unsought results in addition to the traits the farmer seeks [15]. These unintended traits are products of Darwinian unconscious selection.

An example of unconscious selection is that of competitive release in oncology, where a clinician treats a cancer patient in a way that encourages the development of drug resistance. If resistance is secondary, it should be seen, rather than as unconscious, as a product of artificial selection: the oncologist induces resistance by inadvertently selecting for the clones likely to resist the targeted therapy, immunotherapy, chemotherapy, and so forth. Once sensitive cells are eliminated, the resistant cells are able to propagate without the inherent restraint of intratumoral competition.

Like the farmer's artificial selection, the oncologist's unconscious selection works via the same mechanism as ordinary natural selection but should be understood in this context as the unintended result of the regimes the oncologist chooses to treat her patients. Artificial and unconscious selection should therefore be seen as two distinct subsets of evolutionary selection that depend on the knowledge state of the agent, and once the distinction is integrated into clinical oncology, drug resistance will necessarily be viewed as a product of artificial interventions.

In the clinical setting, however, having only understanding of this set of distinctions in evolutionary theory is insufficient for personalizing medicine for cancer patients. Artificial intelligence makes implementation of evolutionary theory in oncology possible: not only can we predict how tumors will respond to targeted therapies that eliminate one part of a cellular population, but we can adjust treatment to reflect expected biological changes. Noted examples in the clinical setting that have reached at least phase 1 study are evolutionary traps, where the biological makeup resulting from artificial selection is specifically targeted by a second-line therapy [16], and adaptive therapy, where competitive release is intentionally unreached to maintain intratumoral competition [17, 18]. In the latter case, which is growing in popularity in mathematical oncology, cure is not the intention; instead, patients can live with a slowly growing tumor for an extended period. Other methods, including treatment vaccination of established infectious risks [19], are of increasing interest – and as the broad field of evolutionary oncology grows and our ability to

pinpoint specific cancer causes and adaptive changes improves, the number of evolutionarily informed treatment options in oncology will expand dramatically.

Future Directions

Dujon et al. recently published a paper highlighting issues raised by leading experts in evolutionary medicine and mathematical oncology in an effort to determine which questions in the ecology and evolution of cancer are the most pressing [20]. Overall, the authors determined 84 questions that must be answered as mathematical oncology moves forward, ranging from how to obtain insights from cancers in wild species to how best to evaluate conceptual and mathematical models of cancer progression. The future, the authors note, will necessarily require interdisciplinary collaboration among evolutionary theorists, research oncologists, statisticians, and engineers for the next major transition in the clinical treatment of cancer. While novel therapeutic strategies, such as adaptive therapy, confirm that evolutionary considerations are integral to improving clinical outcomes, widespread embrace of this new paradigm, grounded in the relationship between artificial intelligence and evolutionary reasoning, is essential for reaching the next adaptive peak on the scientific landscape in cancer treatment.

Infectious Disease

Infectious disease is another field at the crossroads of evolutionary theory and artificial intelligence. The dynamic interplay between pathogen and host environment is a paradigmatic case study for evolutionary theory. Selection pressures, for instance, drive the ominous rise in the prevalence of multidrug-resistant (MDR) organisms. With the massive population sizes and short generation times of bacterial colonies, resistant strains readily emerge when subjected to strong selection pressures from the immune system and antimicrobial drugs, as Darwinian theory accurately predicts.

The European Centre for Disease Prevention and Control (ECDC) maintains the antimicrobial resistance interactive database, EARS-Net, which documents the strong correlation over time between increasing antibiotic use and increasing prevalence of antimicrobial resistance. The evolution of drug resistance is also studied at the individual level, where resistance can arise in the course of individual therapy, such as during treatment for tuberculosis [21] or for chronic diseases such as cystic fibrosis, where recurrent infections are common [22].

Evolutionary theory is also implicated in the emergence of novel human pathogens. These pathogens cross from nonhuman reservoirs to human populations due to ecological and evolutionary factors, and genetic changes, such as neutral drift or adaptive evolution of the pathogen during chains of transmission in humans, lead to the possibility of epidemic disease [23].

Artificial intelligence is indispensable to analyzing and interpreting the explosion of data accumulating in the genome sciences, and it is revolutionizing the field of infectious disease in the process. For instance, Abudahab et al. describe a machine learning model to go beyond the typical phylogenetic modelling of mutations in the core genome to incorporate non-core elements, such as movement of phage, plasmids, and other mobile elements, enriching our understanding of bacterial evolution [24]. As more genomic, proteomic, and transcriptomic data are combined with clinical phenotypes, and with historical information regarding treatment history and healthcare outcomes, machine learning techniques will rapidly improve our ability to monitor and treat infectious diseases.

The purpose of this section is to demonstrate the importance of evolutionary theory to infectious disease and to show how anticipatory, machine learning diagnostics will be essential for further advances in the field, leading the way to next-generation interventions. We discuss the research specifically in regard to drug resistance, drug design, and emerging pathogens and epidemic disease.

Drug Resistance

Starting in the 1940s, the increasingly widespread use of antibiotics in medicine and agriculture has driven the rise in MDR pathogens. The scope of the problem is vast, and the consequences are dire, with the World Health Organization warning that resistance to antimicrobial drugs has reached alarming levels that threaten the achievements of modern medicine [25]. To counter the rise of MDR organisms, machine learning techniques are poised to help provide novel solutions to urgent problems at the patient and population levels.

At the patient level, genomic diagnostics can decrease the time to diagnosis, rapidly identify a pathogen's current resistance profile, and predict its future potential for evolution of drug resistance by shedding light on intrahost resistance evolution [26, 27]. Studies have shown that machine learning methods can accurately and rapidly predict resistance phenotypes from genotype data for HIV protease and reverse transcriptase [28] and for mycobacterial infections [29]. A machine learning-driven analysis of drug resistance in HIV protease, combining protein structure and sequence, demonstrated that resistance tends to be conserved and that there is a secondary selective pressure to become highly resistant over time, while isolates that do not become highly resistant tend to lose resistance [30]. These insights into drug resistance – inspired by evolutionary theory and made possible by artificial intelligence – are crucial to the development of next-generation therapies.

Computational strategies to quantify the genetic potential to develop drug resistance are also being developed, which have the potential to decrease the likelihood of resistant strains emerging [31]. Evolutionary models can inform these methods in order to design and test multi-drug regimens that utilize different strategies to select against specific resistance mechanisms, for instance, using resistance mechanism inhibitors (e.g., beta-lactamases) and selection inversion using suppressive drug interactions, synergy-inducing drug pairs, and collateral sensitivity [27]. Ideally, these strategies can be used to design drug regimens specifically tailored to channel pathogens toward less resistance-prone genotypes. For instance, using an *E. coli* model, Nichol

et al. show how an optimal drug regimen can probabilistically channel the bacterial population through “genotype space” to avoid the emergence of resistance [32]. By rapidly establishing a genotypic diagnosis complete with a current and potential resistance profile, machine learning techniques have the potential to devise a drug regimen specific to an individual case that minimizes the probability of developing resistance at the individual case level, which would be a revolutionary achievement in the field.

At the population level, artificial intelligence has the potential to inform surveillance of resistance trends on both a local and global scale and predict the emergence of resistant mechanisms, allowing for the preemptive implementation of targeted preventative measures. Since bacteria are exposed to antibiotics not only in the human body during therapy but also in veterinary medicine, agriculture, and aquaculture, an inclusive, multidisciplinary approach to infectious disease and drug resistance is necessary to effectively address the problem. Metagenomic and functional metagenomic studies show that bacteria harboring resistance genes are present in all ecosystems, for example, in the soil, the human gut microbiome, and extreme environments, such as permafrost [33]. Therefore, environmental monitoring for drug resistance genes will be an important component of an ideal model for anticipatory diagnostics. Beyond surveillance, advanced computational methods have the potential to determine optimal strategies for combating infection and epidemic disease. Factors such as the timing of antimicrobial therapy, choice of drug regimen, drug cycling, and ecology can be manipulated to optimize patient outcomes and minimize the degree of drug resistance [32, 34]. Continued cross-disciplinary collaboration involving ecoimmunology, disease ecology, and data science is crucial to realizing the potential of artificial intelligence at the epidemiological level.

Drug Design

Antimicrobial drug design and vaccine development are being revolutionized by artificial intelligence. The process by which such computational methods can be utilized is explained well by Park

et al.: “A novel method for *de novo* compound design combines deep learning with reinforcement learning, which estimates the statistical relationship between possible actions and outcomes. Neural networks are trained to generate chemically feasible compounds and predict their chemical properties. Then, using reinforcement learning, the program becomes biased towards compounds with desired physical and biological properties” [35]. Such computational methods will integrate multiomics data to elucidate and anticipate target specificity, molecular effects, and pharmacodynamic, pharmacokinetic, and toxicological characteristics of pharmaceutical compounds. In addition, genotype-phenotype correlations on a large scale will have the power to reveal novel resistance genotypes as potential targets for intervention. Studies in the intrahost evolution of Lassa virus, for instance, found that mutations accumulate in epitopes of viral surface proteins, suggesting selection for immune escape and also potential targets for targeted prevention and intervention [36]. This detailed knowledge of viral genomics could be used to develop antiviral drugs to anticipate and interfere with the known pattern of mutations that result from antimicrobial-induced selection pressure. To wit, machine learning and reverse vaccinology were successfully used to identify six potential vaccine target proteins in the SARS-CoV-2 proteome during the pandemic [37].

Emerging Pathogens

Emerging pathogens are of keen interest today due to the SARS-CoV-2 pandemic. Evolutionary models are crucial to understanding how and when pathogens cross barriers from their natural reservoir to human populations. Machine learning methods can identify common risk factors for emergence and highlight the relative roles of ecology and evolution [23, 38]. For example, a comprehensive study of all known RNA viruses employed machine learning techniques to show that viral virulence can be predicted by ecological traits, including human-to-human transmissibility, transmission routes, tissue tropisms, and host geographic range [39]. Machine learning methods

also increase the efficiency of biosurveillance and outbreak detection, accurately predict the spread of novel pathogens [40, 41], and enable early case detection and tracking, such as for COVID-19 [42]. Researchers have combined machine learning methods with remote sensing data (i.e., data collected by satellite or aircraft sensors), local sensing data (i.e., data measured on site such as rainfall), and data from online social media networks and search engines to track and predict outbreaks of dengue virus, malaria, Zika virus, and influenza with impressive accuracy [43]. Researchers are also predicting viral mutations even before a new strain has emerged. Salama et al. utilized rough set theory and neural networks to develop a model to predict new nucleotide substitutions in the Newcastle virus with an accuracy of about 70% [44]. As the prediction accuracy of artificial intelligence continues to improve, evolutionary theory-informed data science will yield increasingly powerful and innovative tools to combat problems in medicine.

Avoiding the Desynthesis

Anticipating and reacting to phenotypic changes in disease, across oncology and infectious pathogen studies, is essential both for designing individual treatment regimens and for public health planning. The works cited above highlight the enormous potential of combining Darwinian theory with predictive modelling, and research in the health sciences is the most useful when both are used: instead of reacting to change, we can predict it and use such changes to our advantage.

Avoiding an evolutionary desynthesis, where genomic studies become separated from the predictive power of Darwinian theory, will require continued collaboration between and among researchers in evolutionary biology, engineering, genetics, and public health. Continued collaboration – the fruits of which are obvious in the work noted above – will allow healthcare systems and policymakers to adequately prepare, and, hopefully, prevent, the next virus with pandemic potential from disrupting the world as we know it.

References

- Dhawan A, Nichol D, Kinose F, Abazeed ME, Marusyk A, Haura EB, Scott JG. Collateral sensitivity networks reveal evolutionary instability and novel treatment strategies in ALK mutated non-small cell lung cancer. *Sci Rep.* 2017;7:1232. <https://doi.org/10.1038/s41598-017-00791-8>.
- Goodman JR, Ashrafi H. The promising connection between data science and evolutionary theory in oncology. *Front Oncol.* 2020;9:1527. <https://doi.org/10.3389/fonc.2019.01527>.
- Aktipis CA, Kwan VSY, Johnson KA, Neuberg SL, Maley CC. Overlooking evolution: a systematic analysis of cancer relapse and therapeutic resistance research. *PLoS One.* 2011;6:e26100. <https://doi.org/10.1371/journal.pone.0026100>.
- Gatenby RA, Maini PK. Mathematical oncology: cancer summed up. *Nature.* 2003;421:321. <https://doi.org/10.1038/421321a>.
- Anderson ARA, Maini PK. Mathematical oncology. *Bull Math Biol.* 2018;80:945. <https://doi.org/10.1007/s11538-018-0423-5>.
- Wingo PA, Ries LA, Rosenberg HM, Miller DS, Edwards BK. Cancer incidence and mortality, 1973-1995: a report card for the U. S. Cancer. 1998;82(6):1197-1207. [https://doi.org/10.1002/\(sici\)1097-0142\(19980315\)82:6<1197::aid-cncr26>3.0.co;2-0](https://doi.org/10.1002/(sici)1097-0142(19980315)82:6<1197::aid-cncr26>3.0.co;2-0).
- Li X, Blount PL, Vaughan TL, Reid BJ. Application of biomarkers in cancer risk management: evaluation from stochastic clonal evolutionary and dynamic system optimization points of view. *PLoS Comput Biol.* 2011;7:e1001087.
- Kuchenbaecker KB, Hopper JL, Barnes DR, Phillips KA, Mooij TM, Roos-Blom MJ, Jervis S, Van Leeuwen FE, Milne RL, Andrieu N, Goldgar DE, Terry MB, Rookus MA, Easton DF, Antoniou AC. Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. *JAMA.* 2017;317:2402. <https://doi.org/10.1001/jama.2017.7112>.
- Di Fiore R, D'Anneo A, Tesoriere G, Vento R. RB1 in cancer: different mechanisms of RB1 inactivation and alterations of pRb pathway in tumorigenesis. *J Cell Physiol.* 2013. <https://doi.org/10.1002/jcp.24329>.
- Soussi T, Wiman KG. TP53: an oncogene in disguise. *Cell Death Differ.* 2015. <https://doi.org/10.1038/cdd.2015.53>.
- Druker BJ, Tamura S, Buchdunger E, Ohno S, Segal GM, Fanning S, Zimmermann J, Lydon NB. Effects of a selective inhibitor of the Ab1 tyrosine kinase on the growth of Bcr-Ab1 positive cells. *Nat Med.* 1996;2:561. <https://doi.org/10.1038/nm0596-561>.
- Okunade KS. Human papillomavirus and cervical cancer. *J Obstet Gynaecol.* 2020;40:602. <https://doi.org/10.1080/01443615.2019.1634030>.
- Herbein G. The human cytomegalovirus, from oncomodulation to oncogenesis. *Viruses.* 2018;10:408. <https://doi.org/10.3390/v10080408>.

14. Hausen Hz. Is smoking the sole factor in lung cancer development? IASLC 17th world conference on lung cancer. 2016.
15. Heiser CB. Aspects of unconscious selection and the evolution of domesticated plants. *Euphytica*. 1988;37: 77. <https://doi.org/10.1007/BF00037227>.
16. Antonia SJ, Mirza N, Fricke I, Chiappori A, Thompson P, Williams N, Bepler G, Simon G, Janssen W, Lee JH, Menander K, Chada S, Gabrilovich DI. Combination of p53 cancer vaccine with chemotherapy in patients with extensive stage small cell lung cancer. *Clin Cancer Res*. 2006;12:878. <https://doi.org/10.1158/1078-0432.CCR-05-2013>.
17. Gatenby RA, Silva AS, Gillies RJ, Frieden BR. Adaptive therapy. *Cancer Res*. 2009;69:4894. <https://doi.org/10.1158/0008-5472.CAN-08-3658>.
18. Zhang J, Cunningham JJ, Brown JS, Gatenby RA. Integrating evolutionary dynamics into treatment of metastatic castrate-resistant prostate cancer. *Nat Commun*. 2017;8:1816. <https://doi.org/10.1038/s41467-017-01968-5>.
19. Batich KA, Reap EA, Archer GE, Sanchez-Perez L, Nair SK, Schmittling RJ, Norberg P, Xie W, Herndon JE 2nd, Healy P, McLendon RE, Friedman AH, Friedman HS, Bigner D, Vlahovic G, Mitchell DA, Sampson JH. Long-term survival in glioblastoma with cytomegalovirus pp65-targeted vaccination. *Clin Cancer Res*. 2017;23(8):1898–909. <https://doi.org/10.1158/1078-0432.CCR-16-2057>.
20. Dujon AM, Aktipis A, Alix-Panabières C., Amend SR, Boddy AM, Brown JS, Capp J-P, DeGregori J, Ewald P, Gatenby R, Gerlinger M, Giraudau M, Hamede RK, Hansen E, Kareva I, Maley CC, Marusyk A, McGranahan N, Metzger MJ, ... Ujvari B. Identifying key questions in the ecology and evolution of cancer. *Evol Appl*. 2020. <https://doi.org/10.1111/eva.13190>.
21. Matteelli A, Roggi A, Carvalho AC. Extensively drug-resistant tuberculosis: epidemiology and management. *Clin Epidemiol*. 2014;6:111–8.
22. McCaughey G, Diamond P, Elborn JS, McKevitt M, Tunney MM. Resistance development of cystic fibrosis respiratory pathogens when exposed to fosfomycin and tobramycin alone and in combination under aerobic and anaerobic conditions. *PLoS One*. 2013;8:e69763.
23. Antia R, Regoes RR, Koella JC, Bergstrom CT. The role of evolution in the emergence of infectious disease. *Nature*. 2003;426:8–11. <https://doi.org/10.1038/nature02177.1>.
24. Abudahab K, Prada JM, Yang Z, Bentley SD, Croucher NJ, Corander J, Aanensen DM. PANINI: Pangenome neighbour identification for bacterial populations. *Microb Genom*. 2019;5(4):e000220. <https://doi.org/10.1099/mgen.0.000220>.
25. <https://www.who.int/drugresistance/documents/surveillancereport/en/>
26. Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, Earle S, Pankhurst LJ, Anson L, De Cesare M, Piazza P, Votintseva AA, Golubchik T, Wilson DJ, Wyllie DH, Diel R, Niemann S, Feuerriegel S, Kohl TA, ... Iqbal Z. Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. *Nat Commun*. 2015;6. <https://doi.org/10.1038/ncomms10063>.
27. Baym M. Multidrug evolutionary strategies to reverse antibiotic resistance. *Science*. 2016;351(6268): aad3292. <https://doi.org/10.1126/science.aad3292>.
28. Yu X, Weber I, Harrison R. Sparse representation for HIV-1 protease drug resistance. *Proc SIAM Int Conf Data Min*. 2013;2013:342–9.
29. Pandurangan AP, Blundell T. Prediction of impacts of mutations on protein structure and interactions: SDM, a statistical approach, and mCSM, using machine learning. *Protein Sci*. 2020;29(1):247–57. <https://doi.org/10.1002/pro.3774>. Epub 2019 Nov 25.
30. Shah D, Freas C, Weber IT, Harrison RW. Evolution of drug resistance in HIV protease. *BMC Bioinformatics*. 2020;21(18):1–16. <https://doi.org/10.1186/s12859-020-03825-7>.
31. Theys K, Libin P, Van Laethem K, Abecasis AB. An evolutionary-based approach to quantify the genetic barrier to drug resistance in fast-evolving viruses: an application to HIV-1 subtypes and integrase inhibitors. *BioRxiv*. 2019;63(8):1–12. <https://doi.org/10.1101/647297>.
32. Nichol D, et al. Steering evolution with sequential therapy to prevent the emergence of bacterial antibiotic resistance. *PLoS Comput Biol*. 2015;11(9):e1004493. <https://doi.org/10.1371/journal.pcbi.1004493>. eCollection 2015 Sept.
33. Martinez J, et al. What is a resistance gene? Ranking risk in resistomes. *Nat Rev Microbiol*. 2015;13: 116–23.
34. Carroll SP, Jørgensen PS, Kinnison MT, Bergstrom CT, Denison RF, Gluckman P, Smith TB, Strauss SY, Tabashnik BE. Applying evolutionary biology to address global challenges. *Science*. 2014;346(6207): 1245993. <https://doi.org/10.1126/science.1245993>.
35. Park Y, et al. Emergence of new disease: how can artificial intelligence help? *Trends Mol Med*. 2020;26(7):627.
36. Andersen K, et al. Clinical sequencing uncovers origins and evolution of Lassa virus. *Cell*. 2015;162(4): 738–50.
37. Ong E, Wong MU, Huffman A, He Y. COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning. *bioRxiv*. 2020 Mar 21;2020.03.20.000141.
38. Bergstrom C, et al. The ecology and evolution of antibiotic-resistant bacteria. In: *Evolution in health and disease*. Oxford University Press; 2007. p. 125–38.
39. Brierley L, Pedersen AB, Woolhouse MEJ. Tissue tropism and transmission ecology predict virulence of human RNA viruses. *PLoS Biol*. 2019;17(11):1–18. <https://doi.org/10.1371/journal.pbio.3000206>.
40. Chae S, Kwon S, Lee D. Predicting Infectious disease using deep learning and big data. *Int J Environ Res*

- Public Health [Internet]. 2018 Aug [cited 2020 Mar 28];15(8). <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6121625/>
41. Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *Lancet Infect Dis* [Internet]. 2020 Mar 11 [cited 2020 Mar 28];0(0). [https://www.thelancet.com/journals/laninf/article/PIIS1473-3099\(20\)30144-4/abstract](https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30144-4/abstract)
42. Bansal A, Padappayil RP, Garg C, Singal A, Gupta M, Klein A. Utility of artificial intelligence amidst the COVID 19 pandemic: a review. *J Med Syst.* 2020;44(9):19–24. <https://doi.org/10.1007/s10916-020-01617-3>.
43. Schwalbe N, et al. Artificial intelligence and the future of global health. *Lancet.* 2020;395(10236):1579–86.
44. Salama MA, et al. The prediction of virus mutation using neural networks and rough set techniques. *EURASIP J Bioinform Syst Biol.* 2016;10. <https://doi.org/10.1186/s13637-016-0042-0>.



AIM and the Patient's Perspective

24

David Taylor

Contents

Introduction	352
AIM and the Patient's Perspective	352
The Need for Transparency	352
Regulations and Data Sharing	353
The Public's Understanding of AI	354
The Public's View on Data Sharing	355
Weighing the Benefits and Risks	356
Levelling the Playing Field	357
Gaining the Public's Trust	358
Technologies for Trust	358
Conclusion	360
References	361

Abstract

It is still early days for the use of AI in medicine, and so the public discourse centers around perceptions of the technology rather than actual experiences. The public have major concerns about data sharing and privacy and the ability or otherwise of the technology to act independently of doctors and human oversight. Regulations are complex and poorly

understood by the public, and media attention focused on recent breaches of trust and security threats has heightened public concerns around sharing sensitive healthcare data, especially with technology companies. However, when given the opportunity to evaluate specific AI use cases, patients become more accepting of data sharing but still expect to be asked for consent or otherwise made aware of their data being used. Companies using public data need to be transparent about its use. Public bodies providing data need to do so transparently and ensure that a greater public good is achieved and distributed fairly across society. Furthermore, developers need to ensure data privacy by design and to make use of innovative technologies being specifically developed for this

D. Taylor (✉)
The Patients Association, Harrow, UK

RSM Digital Health Council, Royal Society of Medicine,
London, UK

Imperial College London, London, UK
e-mail: david.taylor@imperial.ac.uk

purpose. AI technologists need to enter a meaningful and open dialogue with the public and to maintain this throughout their development and adoption pathways. Only by forming true partnerships with the public, putting patients back where they belong, at the center of the discussion, will the field develop to reach its true potential.

Keywords

PPIE · Patient engagement · Transparency · Regulation · Blockchain · Consent · Patient experience · Data sharing · Public trust · Anonymization

Introduction

As illustrated by this chapter, there are myriad applications of artificial intelligence and machine learning to medicine, often involving the manipulation of patient-derived data to power research, provide early diagnosis, or achieve operational efficiencies. While much of this data processing will be invisible to the public, the impact on all our lives will be enormous and depends, to a significant extent, on the provision of access to sensitive data by patients, through consent or other mechanisms.

Future healthcare systems will run on prolific, multisource, patient-derived data. By being able to access and use this data, medical professionals, researchers, managers, and AI software agents will be empowered to provide services more efficiently, safely, and accurately than today. In an ideal world, patients would choose to allow their latest smart device to stream data into their health records or to grant access to companies and researchers developing new diagnostics or treatments. Entrepreneurs with promising ideas would be able to harness data with patient permission, without the need for cumbersome regulatory blocks and administrative overhead, to help create a thriving ecosystem of software and hardware to guide and monitor therapy.

Currently patients have only limited awareness of the applications of machine learning and of AI

in general to medicine or of the importance of, and need for, data sharing in research and development. Patient attitudes to data sharing are shaped by the public debate, and there is much concern around the sharing or leakage of health information and over who profits from their data.

Public trust in the industry is essential, so to fully realize the AI-enabled health data economy, we need to raise awareness of the benefits and provide the means to deliver and distribute the resulting public goods fairly across society. This will require the full awareness, understanding, and cooperation of the data subjects – the patients.

The current regulations are complex and obscure to the public. Governments can play a role in helping to stimulate public debate and by providing further clarity as the technology develops. But it is incumbent on the industry to operate transparently and collaboratively, with patient safety, privacy, and security always in mind.

AIM and the Patient's Perspective

The Need for Transparency

The development and testing of AI for medical applications such as clinical decision-making or diagnosis requires public health or personal health data. Data needs to be of high quality and of known and well-understood provenance. Sometimes several linked datasets may be needed, holding further personal or sensitive information. To ensure a thriving medical AI ecosystem, data needs to be readily available to both public and private organizations. For this to happen, governments and citizens need adequate regulatory regimes that help to create public understanding and trust. Regional statistical data produced for public health purposes are usually not controversial and in many jurisdictions are now made freely available under open-source directives [1–3]. However, data garnered from individuals, even when aggregated or anonymized, usually require consent or another legal rationale governing their use, and for this, the utmost openness and transparency are needed. *N.B. Data are anonymized in different*

ways, and sometimes they are pseudonymized, with the possibility of linking datasets that refer to the same person with a separate key even though all the identifying fields have been removed.

Regulations and Data Sharing

In many jurisdictions, there are specific provisions that determine the permitted processing of personal data. For example, under the UK General Data Protection Regulation (GDPR), organizations will need both a lawful basis (from a specified list) to process personal health data and an additional condition (from another specified list) because health data is sensitive. The lawful basis could, for example, arise from explicit patient consent, or be permitted for the public good, or in order to fulfil a contractual obligation with the individual concerned (such as the provision of treatment). The extra conditions for sensitive data include explicit patient consent, the provision of healthcare or social care, and use by not-for-profit bodies or for public health purposes. Furthermore, each of these conditions has specific detailed provisions that must be met. Where consent is needed, it must be freely given, informed (with stated and understood purposes), explicit, and limited in time and scope to the stated purposes. The UK's Information Commissioner's Office (ICO) [4] states that "*Consent means offering individuals real choice and control. Genuine consent should put individuals in charge, build trust and engagement, and enhance your reputation.*" Also, that service organizations should "*Avoid making consent to processing a precondition of a service.*"

GDPR applies to organizations across Europe and outside of Europe too. It covers the processing of personal information by any entity that offers goods or services or that monitors individuals who are physically located in the European Union (EU).

In the USA, the Health Insurance Portability and Accountability Act of 1996 (or HIPAA rule) requires appropriate safeguards to protect the privacy of personal health information and sets limits and conditions on the uses and disclosures that

may be made of such information without patient consent. However, currently, HIPAA only applies to traditional entities, such as health insurance plans and healthcare providers, but does not apply to software development and other technology companies that may wish to collect sensitive patient-generated data. In these cases, patient consent is obtained under various complex "terms of service" agreements, which hardly anyone reads before accepting.

Under HIPAA, there are no restrictions on the use or disclosure of anonymized health information. Data can be anonymized either by removing specific personal identifiers listed under HIPAA or after sufficient processing for an expert to certify that the risk level for re-identification is "very small." Echoing the UK ICO's call for further guidance in this complex field, the Center for Open Data Enterprise and the Office of the Chief Technology Officer at the US Department of Health and Human Services (HHS), after roundtable discussions with government, industry, and academia, highlighted the need for further clarification and guidance, including specifically updating HIPAA's rules around anonymization to meet modern demands. They also called for further clarification for software developers on the appropriate use of patient-generated data [5].

UK public bodies such as hospital trusts or GP surgeries are regarded by patients as authoritative and therefore in a position of power over them, but under GDPR, they would need to show that consent to access a patient's sensitive information has been freely given. For this reason, there is a different basis altogether to enable healthcare organizations to provide treatment and to share electronic health records for treatment purposes: patient data is held under a duty of confidence (a provision in common law) that allows healthcare providers to operate on the basis of implied consent to use patient data for the purposes of direct care, without breaching confidentiality. This form of implied consent, however, does not allow further uses of the data, such as in machine learning, so AI organizations need to consider GDPR as well, and this applies even for pseudonymized data.

The interplay between common law and the duties of privacy under GDPR is complex, and the ICO admits that greater clarity is needed. In their ruling on the case of DeepMind and the Royal Free Hospital, the ICO stated [6] “*Greater clarity is needed and we are committed to working with other bodies including the National Data Guardian and Health Research Authority, to improve guidance and support to the sector so that healthcare organisations like NHS Trusts can implement data-driven technology solutions safely and legally.*” Under GDPR, the hospital’s role is that of *data controller*, and a technology company like Google Health (formerly DeepMind) would become a *data processor*, able to operate on the data only to the extent specified contractually. The ICO investigation had been triggered by media reports that the Royal Free Hospital had given DeepMind access to unnecessarily large numbers of personal health records, but the ICO later accepted that this was in fact required during the software’s clinical testing phase to ensure patient safety. At the time that the ICO investigated, they found that the Trust had not fully complied with data protection laws, and this was later rectified contractually. The Royal Free [7] later stated “*The data used to provide the app has always been strictly controlled by the Royal Free and has never been used for commercial purposes or combined with other Google products, services or ads – and never will be.*”

In the USA, a patient brought a claim against Google and the University of Chicago, alleging that the University Medical Center had released confidential medical information to Google without proper de-identification. Over 100,000 records had been shared without removing attendance timestamps. A judge dismissed the case on the grounds that the patient had not suffered losses as a result, but the case highlighted the possibility that tech companies could use unrelated datasets (in this case, timestamps from searches or location data from phones) to re-identify otherwise anonymized records [8, 9].

Despite the official findings, these cases have raised media attention and public fears around personal health record sharing with technology companies, and these perceptions were

heightened by the case of Facebook and Cambridge Analytica [10].

The ICO in the Royal Free case commented that “*Organisations must assure themselves and document how they have taken appropriate steps to mitigate data protection risks beyond contractual obligations and the obligation on Google Health under data protection law, such as audits, reports, and other appropriate measures.*” In other words, openness, transparency, and candor are vitally important for public trust.

The Public’s Understanding of AI

Artificial intelligence is a blanket term that is often misused or misunderstood by the media and in popular culture. Compounded by social media’s potential to spread falsehoods and exaggerations, it is no wonder that public understanding of AI and of its role in society is poor. However, there is little doubt that most people have heard of AI and are likely to have views on its uses and dangers.

In a recent US survey on attitudes toward AI, the median respondent predicted a 54% chance that by 2028 AIs would be able to perform most tasks of economic relevance better than humans, with more people believing this to be harmful to humanity rather than beneficial [11].

There are benefits to effective communication by the media. Stories in the popular press can help the public to fathom the real benefits and risks associated with the technology, its actual pace of development, and its current prevalence, but they can also sensationalize aspects of the technology, so we need to probe deeper to discover what patients and the public think about these topics.

According to surveys in the USA and UK [11, 12], most respondents claimed familiarity with AI, but fewer than 15% believed they had encountered it, and only 2% believed it had any current impact on society. Furthermore, there was confusion over those everyday technologies that were in fact powered by AI. Most respondents correctly identified personal assistants and smart speakers as using AI, but only 35% knew that Google Search did so, and fewer than 30%

realized that Google Translate or Amazon's recommendation engines did so.

The Public's View on Data Sharing

Data, of course, is key to the use of AI in medicine, and personal healthcare records are sensitive, being considered by most to be both personal and private. Naively, patients often believe that healthcare records are owned and should be controlled by the patient, and it is possible to imagine a future in which this becomes true through suitable regulation and the technologies described later in this chapter. But, as we have seen, the current situation is a lot more complex and nuanced. It is revealing that in the UK survey, most people did not realize that personal data could or would be used by AI to perform tasks, and around half of respondents were not comfortable with the thought of their personal information being used by AI to perform tasks for them. Personal and operational data are as vital to medical research as they are to AI product development. But a mixed-method study [13] found that almost two thirds of people were unaware that the UK National Health Service (NHS) granted access to such data for research and development purposes. Participants were engaged in a broad and deep discussion on the value of healthcare data to various stakeholders such as patients, the NHS, industry, charities, and academia. This confirmed a high level of interest in how researchers and corporations would use healthcare data. Most people (almost 75%) believed that the public should be involved in decisions about providing specific organizations access to NHS data, and their main consideration was of course the benefits that would accrue to patients. Over 80% of participants were also concerned to ensure that those benefits would be fairly distributed across the country.

This sensitivity to data sharing for secondary purposes, and in particular for AI research, has been seen in many other studies. In the UK, the NHS is the organization that the public most trusts with their personal data if it leads to an improved service or capabilities [14]. Fifty-five percent of

the public trusted the NHS against 22% trusting government, with even fewer trusting pharmaceutical companies (15%) or Internet companies (8%), and surprisingly only 11% trusted charities (academic institutions were left out of the survey). Over half of those surveyed were positive toward AI, but almost the same number were especially worried about their data privacy when AI was involved. An EU consultation found that 73% of EU citizens want to share health data for research and innovation provided the data is secure and only accessible by authorized parties [15]. This high level of acceptability of data sharing for research has been borne out in other studies covering the UK and Ireland. However, citizens still expected to be asked for consent when their identifiable data was used, and enthusiasm was much lower concerning de-identified data to be used without consent [16]. Reasons for concern around de-identification also emerged in several qualitative studies where participants questioned what would qualify as identifying information, whether de-identification could be achieved effectively, whether it was sufficient for the elimination of consent, and whether there were risks of re-identifying individuals. Those surveyed wanted the security of records to be ensured and called for private profit to be capped and for access to third parties to be vetted. In several studies, participants also indicated their preference to retain control over the data in their own electronic health record with explicit opt-in consent and the right to withdraw at any time, coupled with an ability to tailor their sharing preferences for different purposes.

The situation is more complex than surveys can reveal. Citizens' juries enable the public to make considered, informed judgments about complex matters by providing adequate time and a framework for deliberation in a group discussion. Citizens' juries in England [17] were more accepting of data sharing to both private and public sectors after the discussion process. Many participants accepted commercial gain if public benefit was achieved. Data sharing for efficiency gains was acceptable too, provided that the overall aim was public good rather than profit. And, in practical terms, as of October 2020, fewer than

3% of England's NHS patients had used the national opt-out mechanism to prevent their confidential healthcare records being made available for research or planning purposes.

There is widespread belief among patients that they own their healthcare data, even though there is no substance in law for this. No doubt this contributes to the public's data sharing hesitancy. However, when properly informed, patients and the public accept the need for data to be shared with researchers and commercial organizations. They accept that this is necessary and desirable in order to develop AI products and services for public good, provided that there is full transparency and that the benefits are shared across society.

Weighing the Benefits and Risks

Patient experience is what the process of receiving care feels like for the patient and is one of the most important aspects of healthcare provision, strongly influencing patient recall and quality and safety outcomes. For example, in patients undergoing elective surgery, the aspect of experience most strongly associated with a better health outcome was the level of communication with and trust in their doctor [18]. As AI in medicine matures, we should expect that its impact (if any) on patient experience is what will matter most to patients, while healthcare professionals will tend to concentrate more on outcome measures. For further reading, see [19].

It is indeed early days for the use of AI in medicine, and so the public discourse centers around perceptions of the technology rather than actual experiences. However, we can expect patient concerns to center around data sharing and privacy and the ability or otherwise of AI-enhanced technologies to act independently of human control.

A UK study [20] concentrated solely on machine learning (ML) and found healthcare to be the area accepted to have the greatest potential for this technology and where its social benefits most outweighed the potential risks. Despite a low level of understanding of the term ML or of how it

works, the public nevertheless appreciated that computers could improve diagnostic accuracy and medical decision-making by being able to consider more data than doctors. However, participants stressed the need for human doctors to remain involved, to ensure personal contact for genuine engagement and empathy. There was also concern that machines might make broad generalizations about groups of people, rather than producing a tailored or individual response, and that this could lead to an inaccurate diagnosis, which could have far-reaching repercussions for a patient's treatment options and well-being. Despite this, people readily appreciated that machine decision-making could be more accurate and less error-prone or subject to bias than that of a human. They also understood the benefits of ML for resource allocation and public health and its potential to reduce pressure on public services.

In the UK, media attention has focused on negatives of increasing automation such as cybersecurity breaches (e.g., the WannaCry ransomware attack) and some notorious deals between specific hospital trusts and technology companies (e.g., Royal Free Hospital and DeepMind) and on government mishandling of data sharing practice (e.g., Care.data [21] and the sharing of patient data for immigration control [22]). The WannaCry attack in May 2017 did not specifically target the NHS but nevertheless had widespread impact, causing significant disruption and public attention [23].

In the USA, cybersecurity attention has been focused on the targeted theft by hackers of medical records from health insurance and technology companies. The largest to date was Anthem Insurance Company who suffered an attack by a foreign government in 2014, resulting in the loss of 80 million personal data records [24]. In fact, according to Accenture, one in eight consumers in England and one in four consumers in the USA had their health data stolen from healthcare service providers [25].

The Facebook/Cambridge Analytica scandal broke in 2018 on both sides of the Atlantic. The affair brought to the public's attention the possibility of their personal data being used for unspecified purposes and that it may even have been used to influence their political judgment in

the 2016 US election [26]. This contributed to the Federal Trade Commission's decision to fine Facebook \$5 billion and to impose a structure of accountability, transparency, and oversight.

For AI technology to be effective and universally applicable, there is a need for developers to have increasing access to volumes of high-quality health and lifestyle data, and, as we have seen, this generates a tension between personal privacy and consent. Regulation can go some way to address this tension, but there are specific issues of intellectual property ownership and commercial interests and ethical issues such as ensuring equality of access and lack of bias that, for AI, demand even more transparency, candor, and trust than for other digital technologies.

In written evidence to the UK House of Lords AI Committee, the ICO stressed the need to prepare the public for the more widespread use of AI especially in terms of transparency and effective regulatory oversight [27]. With the possibility of AI-enabled technologies making significant decisions about people, with perhaps little or no human oversight, it becomes clear that data protection rules have become more relevant than ever and if applied effectively can help to protect individuals, mitigate risk, and allow society to reap the benefits of AI technology.

As we know, the use of AI in medicine is growing at pace, and there is clearly a lot of scope for further dialogue with the public, so that society can understand and weigh AI's real benefits and risks when applied to medicine. Only then will patients become empowered to exercise their rights and be able to hold practitioners to their obligations. In this respect, governments can help to build public trust and confidence in the technology as well as explain the risks.

There are further issues that could be explored in AI ethics, bias, explainability, and real-world behaviors that are discussed in the research literature and are beginning to enter public discourse. These issues will no doubt become more prominent and relevant to patients as the field matures.

In the UK, regulators have begun to discuss how they should approach regulatory issues in AI, and in both the UK and Canada, regulatory sandboxes have been set up where patients, providers,

clinicians, suppliers, and regulators can experiment with pilot projects in a "safe space" [28, 29]. In the UK, the principal digital transformation body, NHSX, is attempting an integrated approach across the different regulatory bodies [30]. In the USA, the FDA concentrating on AI in medical devices has spelled out some of their key issues including regulation across the lifecycle, transparency and labelling, algorithmic bias, and change control for devices that are able to learn once deployed and in the field [31].

Levelling the Playing Field

We should be mindful of the need to maximize the value created from data generated using public funds and of the social benefits and efficiencies that can be gained from its more liberal use. This is particularly important where healthcare is provided for free at the point of care, as it is in the UK.

Some data is publicly owned (as it has been paid for by public bodies), while other data is commercial (funded and held privately). The distinction is often blurred (especially in the public mind) as it is often necessary for public data to be labelled with metadata and appropriately enhanced by technology companies in order to render it useful for AI development. Indeed, there is a need for companies to maximize their returns from the significant investment that may be required to extract value from that data.

In the past, a workable commercial strategy has been to achieve lock-in for customers, for example, using a specific platform product such as an electronic health record. Other forms of vendor lock-in can result from overly restrictive or exclusive contracts between healthcare providers and an AI vendor. There is of course a need to ensure a level playing field whenever data is shared with a commercial enterprise. The NHS has now banned local healthcare organizations from entering into exclusive arrangements involving their data, and a new National Centre of Expertise was tasked with ensuring that commercial arrangements provide benefits for patients and the NHS. This new policy ensures that commercial arrangements should not include "conditions limiting any benefits from

being applied at a national level,” and the contractual terms should be “transparent and clearly communicated in order to support public trust and confidence” [32]. Furthermore, it is worth noting that antitrust legislation (e.g., in EU and UK competition law) would ensure that healthcare providers entering into a data sharing arrangement with a commercial AI developer should also be prepared to make the same data available to competing companies on the same or similar terms.

To maximize the utility of healthcare data, healthcare providers (such as NHS Trusts in the UK) need to think about the need to quantify and enhance the quality of the data in their custody. This would entail ensuring its provenance, providing suitable metadata, and taking steps to increase the percentage of completed fields in each record of the database.

Gaining the Public’s Trust

We have seen how important it is to retain the public’s trust, in order to be able to maximize the volume and utility of the data that can be made available for AI development, and for the public to have confidence in the resulting devices and services.

So how can a healthcare provider or AI developer go about gaining public trust in their products and services? DeepMind Health set a good example by holding a series of Patient and Public Involvement and Engagement (PPIE) meetings starting in 2016 [33]. They formed a user group/discussion forum that was active throughout 2017/2018 [34], and this culminated in a Patient User Group event in 2019 after their merger with Google Health, at which time the group was, unfortunately, dissolved. This group went beyond the contribution of patients to individual projects to reflect on what effective patient participation should be and what it could achieve. The discussions engaged a diverse group of enthusiastic citizens from around the UK and aspired to debate issues or concerns about the work of the organization and even for patients to initiate and participate in research themes. Other aspirations were for the company to be open and transparent

about its projects and to foster two-way discussion between the group and the company. This set a minimum standard that healthcare providers and the commissioners of AI should adopt in their own patient collaboration endeavors. It is incumbent on the whole value chain including researchers, developers, providers, regulators, and other public bodies to adopt this approach in order to fully gain public trust in AI. Companies using public data need to be transparent about its use. Public bodies providing data need to do so transparently and ensure that a greater public good is achieved and distributed fairly across society.

So, there are benefits for organizations involved in the development and adoption of AI technologies to set up diverse patient collaboration groups and conduct detailed conversations transparently with them. Patient representatives should be consulted from the outset of research and development, and this should continue through deployment and beyond. No questions raised by the group ought to be out of bounds, not even sensitive commercial arrangements.

Technologies for Trust

Beyond meaningful consultation, collaboration, and discussion, there are some specific new technologies that AI developers should consider. Their use can help to provide patients and the public with the assurances and transparency they demand.

Privacy by Design

It is possible to ensure that personal information is never transmitted beyond the healthcare provider’s firewall, so that it always remains subject to their information governance. Organizations can ensure that personal data is aggregated and that it is then transmitted externally in encrypted and summarized form. This approach to supporting privacy has been used successfully to help researchers conduct population-wide studies, requiring, for example, only patient counts to be transmitted externally, and this can speed the process of designing and recruiting patients to a clinical trial (e.g., the European

project *Electronic Health Records Systems for Clinical Research* [35]).

Of course, many projects still require individual healthcare records in which case it is necessary to remove further information to prevent re-identification. However, even best practice methods for de-identifying records have been called into question as potentially unsafe, and so the practice of simply releasing anonymized datasets to third parties is increasingly questionable [36].

In view of these risks and the possible need for ongoing and informed patient consent, innovative technologies are being trialed to help researchers to securely make use of sensitive healthcare data. For example, several universities and other organizations have developed verifiably secure computing environments [37].

Increasingly AI needs to be trained with large datasets, beyond the confines of an individual organization. Such large, nationwide, or international datasets pose even greater security risks and privacy concerns and present added regulatory complications where national boundaries are crossed. For example, it is technically straightforward for medical equipment manufacturers to collect performance data from devices scattered across the globe, but this presents them with regulatory hurdles. Similarly, regulatory hurdles are faced by pharmaceutical companies engaged in clinical trials for new drugs. They simply are not allowed to directly contact patients who have been involved in the trials, for follow-on research or for longer-term follow-up. Blockchain or distributed ledger technologies are becoming recognized for use as platforms for regulatory compliance. They can establish a trusted verifiable audit trail, to track compliance and streamline access to data.

Another promising field is that of *differential privacy*, which enables data to be processed while guaranteeing that no individual patient's data can be discovered from the dataset. This enables AI models to be trained without effectively sharing anyone's sensitive information [38]. Together with federated computing, which can link disparate organizations in a collaborative learning network where each retains its own data, or with techniques for machine learning from data that always remains encrypted, these technologies can

help ensure privacy by design [39]. Blockchain can also be used to protect patient identities while still allowing companies and researchers to obtain consent and access their data [40]. And *dynamic consent* mechanisms can ensure that this consent is informed and current, giving patients the degree of control that they tell us they want.

Blockchain

Blockchain is a form of distributed ledger technology, popularized by cryptocurrencies such as Bitcoin. This is a secure method of collective bookkeeping via the Internet whereby multiple parties can achieve consensus while each maintains an identical copy of a single database. Instead of feeding data from multiple systems into a centralized information exchange for processing, the multiple systems together maintain a single record of truth using irreversible cryptographic hashes [41].

Patient-controlled healthcare records are extremely suited to this kind of treatment. Imagine that it were possible to compile an up-to-date verifiable real-time ledger comprising a record of every encounter a patient has ever had with a healthcare professional. These encounters would be a series of sequential events which over time involve multiple organizations. Later events often depend on earlier ones. Each transaction would describe the addition of a resource to the official patient record by an authorized person, with the resource itself (the substance of the health record) stored externally with a pointer to its location. Similarly, it would be possible to track and audit lookup events as transactions in the ledger and incorporate the patient's own personal, lifestyle, or physiological records from their wearables. This very much resembles a financial ledger, where blockchain has already proven itself.

One of the other benefits of blockchain technology is that it helps with establishing data provenance and value discovery, both essential to a thriving data economy (e.g., Ocean Protocol [42]). Machine-generated data could be stored with information about the type of equipment used, its calibration status, and a precise timestamp. Patients' own devices would record information into the transaction log and would

be clearly identifiable as such. Algorithms used to process the data would also leave their imprint in the record. As these algorithms become more critical to healthcare delivery, with black box machine learning increasingly a feature of diagnosis and care, the ability to review and manage the entire supply chain of data across the population becomes important. If an error is later found in an algorithm used to make a diagnosis, it then becomes a simple matter to backtrack in order to identify and update everyone whose diagnosis or treatment could be affected.

Federated Computing

Federated computing enables healthcare organizations to pool data for machine learning while protecting both their own intellectual property and their patients' personal data. Sensitive data remain in protected zones within each site, unaltered and uncompromised, where they are transformed to a common data model and further processed for machine learning. Each organization in the network can only interrogate or analyze its own data, and only the results of data analysis are shared externally, hence protecting all parties. The technique can be used for formal clinical trials or for real-world data including observational trials, registry data, population, and operational data. In Europe, this approach has been used to build a federated data network capable of allowing access to the data of more than 100 million EU citizens standardized to a common data model and has been used successfully to provide a rapid response to Covid-19 [43].

Dynamic Consent

Since it is often necessary to process individual personal health records whose de-identification is not a reliable process, explicit patient consent may be the best basis for acquiring data for AI research and development. In addition, the public when given the opportunity have expressed a preference to be able to control their own data, to be able to withdraw consent at any time, and to tailor their sharing preferences for different purposes. Dynamic consent is a mechanism for accomplishing this at a detailed level of granularity. Recently there has been much interest in using blockchain to provide this mechanism without

centralized control, even enabling communication among patient stakeholders to aid with recruitment [44]. This would also help to make the documentation of consent transparent and traceable and foster increased trust and greater patient engagement than current (often paper-based) methods of obtaining consent.

When mature, these privacy-preserving and data-enhancing technologies would shift the balance of power back to society, as patients would ultimately be in control of sharing a comprehensive personal healthcare record or aspects of it with researchers in a transparent, trusted framework, with clinical governance principles verifiably followed. And this would help to shift the balance, from large technology companies and data aggregators toward small businesses and researchers with a great idea who could easily read from, or write to, a patient's record with straightforward authorization or consent. When scaled to entire populations, artificial agents could browse the data and make inferences that would aid in clinical decision-making or to help patients to take pro-active measures to maintain their good health. And this would be accomplished while maintaining the individual's ultimate right to privacy and consent.

Conclusion

AI technology is advancing on a broad front, and the public level of understanding currently misses the richness and diversity of the technology. Public perception is shaped by the media, and discussions tend to focus on specific topics such as data ownership and sharing, cybersecurity, privacy, and whether the technology can make decisions independently of the healthcare practitioner. Governments can help to foster discussion, but in order to establish the level of trust necessary for public support, AI technologists and researchers need to enter into a meaningful and open dialogue with the public and to maintain this throughout the technology development and adoption pathways. Only by forming true partnerships with the public, putting patients back where they belong, right at the center of business, will the field develop to reach its true potential.

References

1. European Legislation on Open Data and the Re-use of Public Sector Information. In: Shaping Europe's Digital Future – European Commission. 2020. <https://ec.europa.eu/digital-single-market/en/european-legislation-reuse-public-sector-information>. Accessed 23 Jan 2021.
2. HHS About HealthData.gov.2020. <https://healthdata.gov/content/about>. Accessed 23 Jan 2021.
3. NHS Digital: Supporting Open Data and Transparency. 2020. <https://digital.nhs.uk/services/supporting-open-data-and-transparency>. Accessed 23 Jan 2021.
4. ICO Consent. <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/consent/>. Accessed 23 Jan 2021.
5. Center for Open Data Enterprise (CODE): Sharing and Utilizing Health Data for AI Applications. 2019. <https://healthdatasharing.org/wp-content/uploads/2020/07/RT1-AI-Summary-Report-FINAL-2020.07.28.pdf>. Accessed 23 Jan 2021.
6. ICO: Royal Free NHS Foundation Trust Update, July 2019. <https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2019/07/royal-free-nhs-foundation-trust-update-july-2019/>. Accessed 23 Jan 2021.
7. Royal Free London NHS Foundation Trust: Information Commissioner's Office (ICO) Investigation | How We Use Patient Information. 2019. <https://www.royalfree.nhs.uk/patients-visitors/how-we-use-patient-information/information-commissioners-office-ico-investigation-into-our-work-with-deepmind/>. Accessed 23 Jan 2021.
8. Wakabayashi D. Google and the University of Chicago are sued over data sharing. The New York Times. 2019. <https://www.nytimes.com/2019/06/26/technology/google-university-chicago-data-sharing-lawsuit.html>. Accessed 26 Jan 2021.
9. Judge Dismisses Data Sharing Lawsuit Against University Of Chicago, Google. FierceHealthcare. 2020. <https://www.fiercehealthcare.com/tech/judge-dismisses-data-sharing-lawsuit-against-university-chicago-google>. Accessed 26 Jan 2021.
10. Cambridge Analytica and Facebook: The Scandal and the Fallout So Far. The New York Times. 2018. <https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html>. Accessed 2 Feb 2021.
11. Zhang B, Dafoe A. Artificial intelligence: American attitudes and trends. SSRN Journal. 2019. <https://doi.org/10.2139/ssrn.3312874>.
12. Holder C, Khurana V, Watts M. Artificial intelligence: public perception, attitude and trust. Bristows LLP. 2018. <https://d1pvkxakgv4jo.cloudfront.net/app/uploads/2019/06/11090555/Artificial-Intelligence-Public-Perception-Attitude-and-Trust.pdf>. Accessed 2 Feb 2021.
13. Hopkins H, Kinsella S, van Mil A, van Mil H. Foundations of fairness: views on uses of NHS patients' data and NHS operational data. Understanding Patient Data. 2020. <https://understandingpatientdata.org.uk/sites/default/files/2020-03/Foundations%20of%20Fairness%20-%20Full%20Research%20Report.pdf>. Accessed 2 Feb 2021.
14. How the UK can win the AI race: What we know, what the public think and where we go from here. KPMG. 2018. <https://home.kpmg/content/dam/kpmg/uk/pdf/2018/09/how-the-uk-can-win-the-artificial-intelligence-ai-race.pdf>. Accessed 2 Feb 2021.
15. European Commission. Consultation synopsis report: transformation health and care in the digital single market. 2018. <https://doi.org/10.2759/18589>.
16. Stockdale J, Cassell J, Ford E. "Giving something back": a systematic review and ethical enquiry into public views on the use of patient data for research in the United Kingdom and the Republic of Ireland. Wellcome Open Res. 2019;3:6. <https://doi.org/10.12688/wellcomeopenres.13531.2>.
17. Tully MP, Hassan L, Oswald M, Ainsworth J. Commercial use of health data – a public "trial" by citizens' jury. Learn Health Syst. 2019. <https://doi.org/10.1002/lrh2.10200>.
18. Black N, Varaganum M, Hutchings A. Relationship between patient reported experience (PREMs) and patient reported outcomes (PROMs) in elective surgery. BMJ Qual Saf. 2014;23:534–42. <https://doi.org/10.1136/bmjqqs-2013-002707>.
19. Kingsley C, Patel S. Patient-reported outcome measures and patient-reported experience measures. BJA Educ. 2017;17:137–44. <https://doi.org/10.1093/bjaed/mkw060>.
20. Ipsos Mori: Public views of machine learning findings from public research and engagement conducted on behalf of the Royal Society. The Royal Society. 2017. <https://royalsociety.org/~media/policy/projects/machine-learning/publications/public-views-of-machine-learning-ipso-mori.pdf>. Accessed 25 Jan 2021.
21. Care.data: How Did It Go So Wrong? BBC News. 2014. <https://www.bbc.co.uk/news/health-26259101>. Accessed 27 Jan 2021.
22. Government Forced to Stop Making NHS Give Patient Data to Immigration Officials. The Independent. 2018. <https://www.independent.co.uk/news/health/nhs-patient-data-immigrant-uk-government-hostile-environment-information-a8343681.html>. Accessed 27 Jan 2021.
23. Ghafur S, Kristensen S, Honeyford K, Martin G, Darzi A, Aylin P. A retrospective impact analysis of the WannaCry cyberattack on the NHS. npj Digit Med. 2019. <https://doi.org/10.1038/s41746-019-0161-6>.
24. McGee MK. A new in-depth analysis of anthem breach. Bank Infosecurity. 2017. <https://www.bankinfosecurity.com/new-in-depth-analysis-anthem-breach-a-9627>. Accessed 27 Jan 2021.
25. 13% Of Consumers In England Have Had Their Healthcare Data Breached. Accenture. 2017. <https://www.accenture.com/gb-en/company-news-release-healthcare-data-breached>. Accessed 27 Jan 2021.
26. Lapowsky I. How Cambridge Analytica sparked the great privacy awakening. Wired. 2019. <https://www.wired.com/story/cambridge-analytica-facebook-privacy-awareness/>. Accessed 27 Jan 2021.

27. Written Evidence – Information Commissioner’s Office. Parliament UK. 2017. <http://data.parliament.uk/writtenEvidence/committeeevidence.svc/evidence-document/artificial-intelligence-committee/artificial-intelligence/written/69635.html>. Accessed 27 Jan 2021.
28. CQC, MHRA: Using machine learning in diagnostic services. 2020. https://www.cqc.org.uk/sites/default/files/20200324%20CQC%20sandbox%20report_machine%20learning%20in%20diagnostic%20services.pdf. Accessed 27 Jan 2021.
29. McCarty M. Health Canada making use of ‘regulatory sandbox’ to address AI. BioWorld MedTech. 2020. <https://www.bioworld.com/articles/498030-health-canada-making-use-of-regulatory-sandbox-to-address-ai>. Accessed 27 Jan 2021.
30. Gould M. Regulating AI in health and care – technology in the NHS. 2020. <https://healthtech.blog.gov.uk/2020/02/12/regulating-ai-in-health-and-care/>. Accessed 27 Jan 2021.
31. FDA: Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. 2021. <https://www.fda.gov/media/145022/download>. Accessed 27 Jan 2021.
32. NHSX. To oversee data-sharing agreements under new DHSC guidance. Digital Health. 2019. <https://www.digitalhealth.net/2019/07/nhsx-to-oversee-data-sharing-agreements-under-new-dhsc-guidance/>. Accessed 27 Jan 2021.
33. Healthwatch Camden: DeepMind Health Patient and Public Engagement Event – 20/09/2016. https://healthwatchcamden.co.uk/sites/default/files/pieceeventssummary_1.pdf. Accessed 27 Jan 2021.
34. King D. Collaborating with patients for better outcomes. Deepmind. 2017. <https://deepmind.com/blog/article/collaborating-with-patients>. Accessed 27 Jan 2021.
35. IMI Innovative Medicines Initiative: Electronic health records systems for clinical research. 2016. <https://www.imi.europa.eu/projects-results/project-factsheets/ehr4cr>. Accessed 3 Feb 2021.
36. Rocher L, Hendrickx JM, de Montjoye Y-A. Estimating the success of re-identifications in incomplete datasets using generative models. Nat Commun. 2019. <https://doi.org/10.1038/s41467-019-10933-3>.
37. Data Safe Havens in the Cloud. The Alan Turing Institute. 2020. <https://www.turing.ac.uk/research/research-projects/data-safe-havens-cloud>. Accessed 3 Feb 2021.
38. Zhu T, Ye D, Wang W, Zhou W, Yu P. More than privacy: applying differential privacy in key areas of artificial intelligence. IEEE Trans Knowl Data Eng 1–1. 2021. <https://doi.org/10.1109/TKDE.2020.3014246>.
39. Kaassis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell. 2020;2: 305–11. <https://doi.org/10.1038/s42256-020-0186-1>.
40. Marchal B. Proposal of a new privacy safe approach to data access in rare disease. 2020. <https://youtu.be/qZxITL7s3uk>. Accessed 27 Jan 2021.
41. Wolpert M. Distributed ledger technology: beyond block chain. A report by the UK Government Chief Scientific Advisor. 2015. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/492972/gs-16-1-distributed-ledger-technology.pdf. Accessed 27 Jan 2021.
42. Roche Diagnostics and Ocean Protocol Partner to Improve Care for Heart Disease Patients Through safe and secure data sharing. Medium. 2019. <https://blog.oceanprotocol.com/roche-diagnostics-and-ocean-protocol-partner-to-improve-care-for-heart-disease-patients-through-71cf3e678dc>. Accessed 27 Jan 2021.
43. European Health Data & Evidence Network: Vision, Mission & Objectives. 2020. <https://www.ehdenn.eu/vision-and-mission/>. Accessed 3 Feb 2021.
44. Porsdam Mann S, Savulescu J, Ravaud P, Benchooufi M. Blockchain, consent and prosent for medical research. J Med Ethics. 2020. medethics-2019-105963. <https://doi.org/10.1136/medethics-2019-105963>.



AIM and Hackathon Events

25

Ayomide Owoyemi and Wuraola Oyewusi

Contents

Introduction	364
Organizing AIM Hackathons	365
Proposed Metrics for Measuring the Success of an AIM Hackathon	367
Limitations of the Hackathon Approach	368
Recommendations for Organizing Better Hackathons	368
References	368

Abstract

Hackathons are leveraged to crowdsource innovation. Hackathons have become an accepted approach to problem-solving and product development in the technology and healthcare sectors. Hackathons are used by companies, organizations, and research teams either internally or in open formats for different purposes. For AIM, hackathons could serve different purposes for different entities which can range from the discovery/development of new products, identification of innovative approaches, innovative application of existing resources or tools by an organization, and collaborative and inclusive design. Measuring the

success will vary and be relative to the set goals of the hackathon, but it is important that organizers come up with open practical questions that can enable them to gauge the effectiveness and usefulness of AIM hackathons. AIM hackathons can be better organized if organizers leverage on existing resources and knowledge, involve more experienced professionals as participants to benefit from their experience and insight, and plan strategically for follow-up to the event for winning teams or solutions.

Keywords

Artificial Intelligence in medicine · AIM · Innovation · Collaboration · Medical hackathon · Multidisciplinary · Community · Team

A. Owoyemi (✉)

University of Illinois at Chicago, Chicago, IL, USA

W. Oyewusi

Research and Innovation, Data Science Nigeria, Lagos,
Nigeria

e-mail: wuraola@datasciencenigeria.ai

Introduction

Hackathon is an intense, short, collaborative event focused on creating innovative solutions for pressing problems. It essentially combines the words “hack” and “marathon” to imply a sense of an intense and continuous event targeted at solving a specific problem [1].

Silver et al. have gone further to specifically coin a definition for healthcare hackathon which they define as

A competitive event (live or virtual) that has three specific goals—accelerating the innovation of medical solutions, improving the design in the beginning stages, and supporting educational training for all participants—and aims to accomplish them by focusing on a specific problem (pain point), bringing together in an open innovation format (internal and external resources) an interdisciplinary group of individuals (hackers) that include, but are not limited to, physicians and other healthcare professionals, data scientists, engineers, user interface designers, business professionals, students and other stakeholders who work in teams and follow a process to develop initial prototypes, pitch them to a panel of judges experienced in innovation and quickly alter them according to the feedback (pivoting). [2]

According to Ramadi and Nguyen [3], hackathons usually try to achieve several goals which are:

- To contextualize an impactful problem to a specific real-life scenario
- To convene cognitively, geographically, and socially diverse teams to tackle the problem
- To assemble an equally diverse set of domain and process expert mentors to provide accelerated customer discovery and subject matter expertise
- To optimize the design for accelerated development and implementation

Hackathons have become an accepted approach to problem-solving and product development in the technology and healthcare sectors. Hackathons are used by companies, organizations, and research teams either internally or in open formats for different purposes [4]. Different companies organize hackathons either internally

or externally to find innovative ways to solve the extant problems the companies face, while others organize them to create a platform for the involvement of multidisciplinary and diverse individuals to collaborate on finding the required solutions.

Hackathons could serve different purposes for different entities which can range from the discovery/development of new products, identification of innovative approaches, innovative application of existing resources or tools by an organization, collaborative and inclusive design, etc. These usually depend on the problem being targeted or the entity organizing the event. Hackathons create an avenue for interdisciplinary and interprofessional collaboration to solve health problems. This is more pronounced in cases that involve the development, iteration, and application of technology which domain experts might not be easily familiar with or understand the underlying principles.

Another often-overlooked goal and likely impact of hackathons is their role as educational avenues for students and professionals alike. They are potentially significant learning experiences that can help promote technical skills, thinking skills, and mastery of specific issues. They also help introduce and create professional and social networks that can be useful to these students and professionals in the long term [5].

Google, Facebook, and Microsoft usually organize internal hackathons for their employees to create new products or features [6]. The Massachusetts Institute of Technology (MIT) has an MIT Hacking Medicine community that has facilitated over 80 hackathons and healthcare design thinking events over the past 5 years. Its model is built on the concept of integrating collaboration and interdisciplinary gathering around innovation techniques. It has successfully launched about 20 active companies that have raised over \$120 million dollars [7].

Hackathons are also gaining grounds in developing countries. For example, there is a dedicated platform designed for competitions and hackathons, Zindi, established in 2018 and based in Cape Town, South Africa. Zindi describes itself as the first data science competition platform in Africa. Zindi hosts an entire data science ecosystem of scientists, engineers, academics,

companies, NGOs, governments, and institutions focused on solving Africa's most pressing problems; this includes health-related problems that can be solved by data and AI. Zindi works with companies, nonprofit organizations, and government institutions to develop, curate, and prepare data-driven challenges [8].

In July 2020, AI Commons collaborated with Data Science Nigeria for the AI Commons Health & Wellbeing Hackathon which was an online competition to develop groundbreaking and innovative working prototypes, models, or solutions driven by Artificial Intelligence that can solve identified local health problems or improve identified existing health technology solutions in Nigeria. The project aimed to increase/improve the accessibility, reproducibility, contextualization, and enhancement of Artificial Intelligence solutions globally and especially in emerging markets. The core participants were AI experts, health workers, academic researchers, health-tech organizations, students, and any other interested party [9].

The relevance of hackathon to solving problems and innovations in medicine became more prominent when COVID-19 became a pandemic, and there was a high demand for accelerated solutions that could be used in solving the public health emergency created by the pandemic. The medical community rose to the occasion and collaborated through different global hackathons which produced different solutions that were applied to varying degrees of success in managing the pandemic [3].

According to DePasse et al. [10], the significance of hackathons revolves around three core principles:

- A problem-based approach which harnesses the perspective and experiences of clinicians, patients, and other likely users in approaching the possible solution.
- Harnessing of diversity of thought, background, and perspective across different disciplines, professions, and experiences creating a platform that can help to shrink the innovation timeline with opportunities for rapid feedback and better productivity.

- Challenge of existing paradigms and easy pivoting. This is based on the possibility of rapid feedback that ensures that developed concepts or products are meeting the actual needs and problems.

These principles are very relevant when it comes to the application or implementation of AI/ML as tools in healthcare because of extant limitations in accessing the required data and resources for developing these tools due to privacy reasons and access to required expertise for both data scientists and domain experts. Also, medical professionals and data scientists and engineers rarely interact, and hackathons can help bridge this gap and create an atmosphere for collaboration and sharing of ideas and perspective. Hackathons also help ensure that use cases are best developed from the perspective of the end users.

Organizing AIM Hackathons

According to Nolte et al. [4], there are 12 key decisions that must be considered to ensure a successful hackathon, but for the purpose of this text, we will categorize these decisions under 10 contexts which are most relevant to AIM hackathons:

Goal and theme: There should be a clear setting of the goals of the event. The feasibility and attainability of these goals should be well considered to prevent disappointing outcomes from the event. These goals can range from the illustration of a concept to production of something useful, learning experience for participants, or connection of professionals. It is also important that these goals are clearly communicated to prospective participants to avoid possible mismatch of expectations. The theme of the hackathon is usually developed in conjunction with the goals of the event. This theme is best arrived at after discussions with relevant stakeholders and experts in that aspect by the organizers. Themes can be general or specific. It is usually much better for hackathons if the

selected themes are specific as they ensure that the work serves a specific outcome; however, specific themes can be limiting for participants especially when it comes to their personal interests or how much they can explore. Organizers will have to consider these tradeoffs based on their set goals.

Format: Organizers need to decide in the designing phase if they want a collaborative or competitive hackathon; this will involve the likely introduction of incentives to stiffen the competition if necessary. Hackathons are usually organized around a commonality, which might involve the development of specific aspects of an existing project, expansion of its capability, introduction of new features to it, or networking for the participants involved. Competitive ones focus on creation of innovations. For competitive events, winners can be decided by a jury or through popular votes. The use of a jury of varied experts might be more appropriate for AIM events.

Stakeholder involvement: The implementation of AIM is a multidisciplinary endeavor, and it is important that stakeholders from all the involved disciplines are involved in the planning processes for the hackathon. It is important that these groups of stakeholders are involved in the organization, execution, and follow-up of the hackathon. Their involvement will be very integral to the real-world deployment and sustainability of the outcomes of the event. These stakeholders could act as sponsors, speakers, coaches, resource persons, mentors, or judges for the event.

Participants: The targeted participant and mode of recruitment for the hackathon will vary based on the goal, theme, and format being planned. The recruitment could also be open or restricted to a closed community. These will affect the period of recruitment and the characteristics of the target participants. Some events might be open to anyone who might be interested; some would want people with a specific skill, technical expertise, and experience in specific domains.

Ideation: The structure of this aspect of the event has a bearing on how interesting and attainable

the ideas that will be worked on will be. Some events usually organize an ideation session before the event to come up with a cluster of ideas that will be eventually worked on at the main event, while others just allow the participants to ideate freely based on the theme. There are trade-offs for both approaches; allowing participants to develop their own ideas and work on them might be more motivating than limiting what can be worked on. There is also the trade-off of hacking time and ideation time, as hackathons are usually of short duration; it might not be helpful in some cases to have ideation during the main event.

Team formation: The strategy for team formation is another important decision that organizers must make in their plans. Teams can be formed by either open selection in which participants select the groups and roles they want to play themselves or assignment where participants are assigned to projects and roles by a mediator or a hybrid model that allows the participants to prioritize projects based on interest and expertise, after which they are assigned by a mediator based on the stated preferences. These teams can either be formed before or at the hackathon; the preference will depend on the goal and theme of the event. A learning and networking event may be best served by a model that allows people to cluster at the event itself. The multidisciplinary nature of hackathons should be leveraged for data interpretation, e.g., participants with medical background interpret and bring context into medical data, and those with other skills such as machine learning algorithms, data wrangling, etc. solve from their areas of expertise.

Resources: One of the most important resources at AIM hackathons is data. One of the barriers to leveraging AI for medicine is data confidentiality and understanding. Questions and concerns around data privacy are valid, and the nature of medical-related data increases these concerns. Data de-identification can go a long way in solving this for hackathons when possible; also the use of synthetic but similar datasets can reduce real data exposure. Organizers can also explore federated learning

systems; federated learning allows the sensitive patient data to stay either in local institutions or with individual consumers without going out during the federated model learning process, which effectively protects the patient privacy [11].

Participants in AIM hackathons should also be required to sign nondisclosure agreements; and role-based access should be activated as appropriate, such that participants only have access to what they need. Organizers of AIM hackathons who can explore the use of virtual machines should also consider this, so participants only have access to data for a limited time. A peculiarity of AI in medicine that planners should prepare well for is the length of training time for machine learning/deep learning models and other needs like Graphics Processing Units that hasten model training.

Agenda: It is important that organizers have a clear agenda for the event as this will also help to manage expectations and allow participants to adequately plan for the event. It is important to make this agenda available to everyone involved a few days prior to the event. It is advisable that extra-hacking activities are limited so that participants can have more time to work on their projects. A significant activity that might be helpful is “checkpoint.” This involves teams reporting their progress, discussing challenges they are facing, and outlining the plans for the time ahead. Checkpoints will be an opportunity for teams to get additional support and insights on what they are working on.

Mentoring: A common strategy is to provide mentors on demand for the teams; this is usually employed when teams are working on their own projects and have the necessary skills required to develop their ideas. Some events use dedicated mentors who are specifically assigned to individual teams or give teams the opportunities to source their own mentors. The most important thing is the accessibility of the mentors for the participants. These mentors can be recruited through an open call which usually happens long before the main event, or they can be handpicked by the organizers from

data science, engineering, medicine, public health, or the specialties and disciplines that are relevant to the theme of the event. In some cases, a team of mentors might be important, e.g., a team made up of mentors with technical skills and domain expertise as this might be more helpful to the teams depending on the level of expertise of recruited participants.

Continuity framework: Depending on the goals of the hackathon, it is important that adequate plans are made for what will happen after the event. This should be planned for before the event and well communicated to the participants when the publicity for the event is being shared. These continuity plans could involve the creation of a community, transformation of projects to startups, connection of solution with support and funding to take it to market, etc. This plan will also help in informing which stakeholders might be relevant to involve in the plan and execution of the hackathon. It is possible that participants might have different plans for continuity but adequately conveying that prior to the event will help manage expectations and planning for prospective participants.

Proposed Metrics for Measuring the Success of an AIM Hackathon

Measuring success of a hackathon will vary and be relative to the set goals of the hackathon, but it is important that organizers come up with open practical questions that can enable them to gauge the effectiveness and usefulness of an AIM hackathon. These could include:

- Were the predefined goals met?
- Did the event run as planned? (Assessments using pre- and post-surveys for all participants)
- Was the judgment of the winning solution fair and in alignment with previously outlined criteria?
- Is the winning solution practical in context? Does it work for the end user (patients, health professionals)?
- What progress did the winning team(s) and solution(s) make thereafter?

Limitations of the Hackathon Approach

Many problems require continuous and long-term iterations in design, implementation, and testing; therefore, hackathons would not suffice in proposing the most appropriate solution to such problems. Hackathons are not intended to be a standalone innovation event or process but rather as a part of an innovation continuum that can help accelerate and improve future solutions [2].

Organizers of AIM hackathons or any hackathons at all attempt to bring together an interdisciplinary team, but the perfect mix is not guaranteed; sometimes important professionals are missing, e.g., an AIM hackathon without medical doctors or AI engineers applying to be part of the team. A way to mitigate this is to include instructions on professionals that should be in a team at the call for participation if the team is self-coordinating; if the decision on team formation is made by the organizers, then they may need to involve important professionals that may be missing in certain groups.

There is also the question of who owns the intellectual property (IP) of innovation during the hackathon. If the participants (who probably do not have an existing relationship) own the IP, how do they share the benefit? Many hackathon organizers solve this by making the innovation open source or taking up the IP. It is suggested that these terms are clear in the call for participation.

Recommendations for Organizing Better Hackathons

Leverage Existing Resources: It is tempting to assume a concept is new, but hackathons have been well explored and documented. It is suggested that organizers do not reinvent the wheel, research, and adapt to what works in your system. This also includes using existing competition platform's infrastructure rather than setting up new ones from scratch.

Involvement: Most hackathon participants are usually drawn from pools of young and early career individuals, while experienced professionals

are mostly used as mentors or coaches. It will be more helpful for the quality of outputs of these events to have these groups of experienced professionals take part as participants. AIM hackathons would benefit from the experience and insight that these professionals offer and involve them in the development of technology solutions.

Continuity: AIM hackathons are meant to be catalytic processes. The follow-up to the event for winning teams or solutions should be well planned during the design process to ensure the outcomes of the event can translate more meaningfully for the targeted problems.

References

- Celi LA, Ippolito A, Montgomery RA, Moses C, Stone DJ. Crowdsourcing knowledge discovery and innovations in medicine. *J Med Internet Res* [Internet]. 2014;16(9). <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4180345/> [cited 1 Mar 2021].
- Silver JK, Binder DS, Zubcevik N, Zafonte RD. Healthcare Hackathons provide educational and innovation opportunities: a case study and best practice recommendations. *J Med Syst*. 2016;40(7):177.
- Ramadi KB, Nguyen FT. Rapid crowdsourced innovation for COVID-19 response and economic growth. *npj Digital Medicine*. 2021;4(1):1–5.
- Nolte A, Pe-Than EPP, Affia AO, Chaihirunkarn C, Filippova A, Kalyanasundaram A, et al. How to organize a hackathon – a planning kit. *arXiv:200808025 [cs]* [Internet]. 2020. <http://arxiv.org/abs/2008.08025> [cited 7 Feb 2021].
- Lyndon MP, Cassidy MP, Celi LA, Hendrik L, Kim YJ, Gomez N, et al. Hacking Hackathons: preparing the next generation for the multidisciplinary world of healthcare technology. *Int J Med Inform*. 2018;112: 1–5.
- The complete guide to organizing a successful hackathon | HackerEarth Resources | [Internet]. Innovation Management Resources. 2017. <https://www.hackerearth.com/community-hackathons/resources/e-books/guide-to-organize-hackathon/> [cited 1 Mar 2021].
- Lee C. Health hackathons. *Int J Infect Dis*. 2016;53:7.
- Zindi rallies Africa's data scientists to crowd-solve local problems [Internet]. TechCrunch. <http://social.techcrunch.com/2019/08/09/zindi-rallies-africas-data-scientists-to-crowd-solve-local-problems/> [cited 24 Nov 2019].
- Data Science Nigeria and AI commons bring Life360. A proof of concept of a new methodology of developing Artificial Intelligence solutions that allows anyone, anywhere to benefit from the possibilities that AI can

- provide. [Internet]. Data Science Nigeria. <http://www.datasciencenigeria.org/ai-commons-hackathon/> [cited 1 Mar 2021].
10. DePasse JW, Carroll R, Ippolito A, Yost A, Santorino D, Chu Z, et al. Less noise, more hacking: how to deploy principles from MIT's hacking medicine to accelerate health care. *Int J Technol Assess Health Care.* 2014;30(3):260–4.
11. Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated learning for healthcare informatics. *J Healthc Inform Res.* 2021;5(1):1–19.

Further Reading

MIT's Hacking Medicine's Health Hackathon Handbook is a comprehensive resource for health hackathons; it covers the range of decisions and

important features of health hackathons; the content is relevant for AIM. https://docs.google.com/viewer?url=http://admin.mithackmed.com/wp-content/uploads/2018/03/Health_Hackathon_Handbook.pdf&hl=en_US

For AIM in Africa, Data Science Nigeria and AI Commons also made the process and content for the AI Commons Health & Wellbeing Hackathon openly available online. <https://www.datasciencenigeria.org/ai-commons-hackathon/>

Problem and Solution Documentation Template for Machine Learning Competitions to Enhance Explainability, Reproducibility, and Collaboration Between Stakeholders https://www.researchgate.net/publication/346676264_Problem_and_Solution_Documentation_Template_for_Machine_Learning_Competitions_to_Enhance_Explainability_Reproducibility_and_Collaboration_Between_Stakeholders



AIM, Philosophy, and Ethics

26

Stephen Rainey, Yasemin J. Erden, and Anais Resseguier

Contents

Introduction	372
Promises of AI in Medicine	372
AI and Medical Epistemology: A Changing Paradigm	373
Data	373
Data-Utopianism	374
Data Curation and Use	375
AI and Medical Epistemology: Limits, Risks, and Biases	376
Human Biases and Prejudices: Language and Interpretation	377
Computational Biases: Programming and Algorithms	378
AI and Medical Ethics	379
The Patient-Doctor Relationship	380
The Medical Profession in the Era of Digital Capitalism	381
Conclusion and Recommendations	382
References	382

Abstract

This chapter explores AI through a philosophical and ethical lens. This includes an examination of how AI impacts on medicine in terms

of uses and promises, limitations, and risks, as well as key questions to consider. While AI offers scope for complex and large-scale data processing, with the promise of an increase in efficiency and precision, some central limitations need to be highlighted. The use of AI also brings some pertinent and predictable, as well as unpredictable risks, such as those due to biases. Also considered is what may be lost where AI replaces established processes, not least those relational and interpersonal aspects that are central to healthcare. By covering these and related issues, this chapter offers ways to evaluate, and also balance, key benefits and

S. Rainey (✉)
University of Oxford, Oxford, UK
e-mail: stephen.rainey@philosophy.ox.ac.uk

Y. J. Erden
University of Twente, Enschede, The Netherlands
e-mail: y.j.erden@utwente.nl

A. Resseguier
Trilateral Research, Waterford, Ireland
e-mail: anais.resseguier@trilateralresearch.com

risks arising from the application of AI to the medical sector.

Keywords

Philosophy · Artificial intelligence · Medicine · Care · Bias · Data · Medical epistemology · Medical ethics · Algorithms

Introduction

This chapter examines AI through a philosophical and ethical lens; it explores some fundamental impacts on medicine in terms of promises, limitations, risks and key questions. On the one hand, AI has the capacity to process huge amounts of data on medical issues. This may increase efficiency and precision for diagnosis and treatment and offer progress in medical research. On the other hand, the possibility emerges of a drift toward over optimistic techno-solutionism in medicine and healthcare. This includes the application of problem-solving approaches which are apt for data-centered practice, but which are not obviously improvements when it comes to other dimensions of medicine, especially its relational aspect. Whereas medicine in general has a clear anchoring in scientific accounting for the biological body, it also has responsibilities in recognizing subjective, interpersonal, sociopolitical, and historical realities about health and illness. The inclusion of AI in medical practice raises interesting and challenging philosophical, ethical, as well as practical questions. In exploring these topics related questions about asymmetric interpersonal relations in medicine, and on topics of personal identity as they relate to AI and medicine will be considered.

This chapter begins by considering the promises of AI (section “[Promises of AI in Medicine](#)”), but gives only limited time to this endeavor since many of the chapters in this volume offer more detailed examples of how AI can fruitfully be used in medicine. The primary aims in this chapter therefore concern on the one hand, AI in relation to philosophy, and more specifically in terms of medical epistemology (section “[AI and Medical Epistemology: A Changing Paradigm](#)”), and on

the other, in relation to medical ethics (section “[AI and Medical Ethics](#)”). The argument is that the epistemological issues are at the core of many of the ethical issues that follow, and so this earlier section is necessarily longer and more detailed. The suggestion is that understanding the limitations of AI in terms of issues arising from data and from bias will enable the reader to more easily anticipate and assess the other ethical issues that are outlined in section “[AI and Medical Ethics](#)”. Many of these also overlap with, or build on, those earlier issues and so can be explained more concisely by that stage.

Before beginning, it is important to note that AI and ethics is a vast area, and there is plenty of literature already written on this topic generally [1], as well as in terms of AI as it pertains to medicine, healthcare and clinical settings. See for instance recent work on whether AI and medical principles can align [2], on the application of AI to specific areas of medicine, including intensive care [3] and on the use of AI in surgery [4]. Meanwhile there are also texts that offer high level mapping, thereby outlining general ethical issues as they relate to AI in healthcare [5]. Mapping offers a useful general picture, while detailed accounts on single applications offer scope for focused analysis of specific issues. However, the aim in this chapter is to offer a practical, middle ground that introduces the reader to a few central topics that are core to thinking philosophically and ethically about AI in medicine. By restricting the focus in this way there can be some detailed discussion of each, while not precluding the importance of other ethical issues relevant to these topics. The reader can therefore use this chapter as an introduction to some foundational issues, and a way to begin the important work of ethical and philosophical analysis of AI as it pertains to medicine. It should thereby be considered a stepping stone to further reading and analysis.

Promises of AI in Medicine

AI applications in medicine may be claimed to be more objective than standard, human-based approaches. Some of the methods employing AI might appear able, to some extent, to bypass

clinician bias. Where, for instance, a human may see a person and make snap judgements or harbor implicit biases about them, AI “sees” only data (this apparent potentiality is explored more critically later in this chapter). This hope for objectivity would be a boost for justice in healthcare. Certainly, it may be true in cases whereby AI is employed to assist in the managing or assessment of data and/or in ensuring that approaches to clinical procedures and processes are consistent and rigorous. While there may be these improvements brought by AI in medicine in terms of justice – through more equitable access or distribution of resources – they may alternatively be seen simply as efficiency boosts for the practical implementation of medical decision-making. Where AI can take over more laborious tasks that otherwise tie up clinicians’ time, this can free those clinicians to do more valuable work.

A large amount of what AI can do in medicine centers on imaging, given the aptitude for pattern recognition in AI applications. This can be seen in radiology, pathology, ophthalmology, and dermatology [6]. In these areas, AI can be used to detect fractures from scanning x-ray images, or potential eye or skin problems from processing photographs of retinas and skin. In this way, the system can assist clinicians to prioritize their work. Besides imaging, an emerging and potentially very powerful application comes from classifying text. Clinicians make detailed notes as they work, including their observations and practices. AI can be set loose on amassed clinical note data, and spot common findings, approaches, mistakes, and inefficiencies [6]. This processing can then lead to recommendations that would allow protocols to be fine-tuned based on analysis of many instances of clinical notes. AI could do this classification in an unsupervised way, freeing up time for clinicians.

Even where the promises of objectivity and efficiency are assumed, a further query arises. *How sure can we be that AI can deliver on those promises?* It isn’t automatically clear that more and more data applied to individual cases is necessarily better than careful attention to the individual specifics. Imaging and clinical note consolidation are valuable applications, but as

explored below it would be prudent not to expect an AI replacement for clinical expertise in general, given much can be lost as well as gained. One risk to be avoided is that of imbuing AI with an “enchanted” status [7], which would serve to close down scrutiny of those applications. Advances in AI technology don’t inevitably lead to improvements in applications. They may represent changes to applications that require evaluation on their own terms.

AI and Medical Epistemology: A Changing Paradigm

Data

Doctors have always collected data (observations, case notes, research findings) in order to come to diagnosis. Typically, these data have been handled so as to inform medical theories, bolstering clinical insights through broad observational regularities. One hope for the future of medicine includes the idea that AI can be used on data in a much more directly instrumental way in order to boost the efficiency, predictive power, and ultimately the effectiveness of medicine. Such a data-centric approach – where research, diagnosis, and treatment primarily derive from the processing of digitized health data, rather than patient observation – is thought by some to permit a computational approach to medicine. This would diminish the role of medical theories, and expertise, replaced by exploratory data science [8].

The promise of data in general, and big health data specifically, is that it can represent vast arrays of knowledge based on samples, processed in various ways, without guiding theoretical knowledge required. The data-centric approach surpasses limitations present in case study or cohort analyses by aggregating wide ranges of quantitative and qualitative observations. This provides scales not available by other means. These mass aggregates of data can be transformed into new knowledge, through applying statistical transformations and pattern analyses. At its simplest, the idea is that where there are patterns in data, there can be reasons to explain those patterns [9]. These

patterns would be taken to signify relationships among arrays of observations. Doubtless, owing to the way in which humans are primed to spot patterns anyway – like pareidolia and seeing faces in clouds – such things will be seen frequently. But there will be a set of patterns among datasets that are not mere coincidence. By exploring data closely, these can be shifted from the burgeoning whole.

Patterns found within data that aren't mere coincidence will have some other causal explanation. This means that the patterns in the data, representing structures implicit among the observations of some phenomena, will reveal causal structures among those phenomena. The relationships among data such that a pattern emerges are thought to reveal fundamental information about whatever the researcher is investigating in this way. And in finding new causal structures among phenomena by examining patterns in data, the formulation of new insights into those phenomena are enabled by examining the data. This in turn can be used as a basis for new practical approaches to the field. This has application in medicine where medical and other data can be combined and from the whole predictions made to explain patterns. The patterns in data are then expected to relate to causal factors in diagnosis, prognosis, health outcomes, treatment, and so on.

For example, demographic information might be aggregated into a large dataset, along with clinical data, genetic information, observations from specimen biological samples, and other meta-data [10]. By turning to data analysis, relationships among this trove of heterogeneous material might serve to yield patterns that suggest underlying structures among lifestyle, medical history, illness, and health outcomes [11]. This in turn can offer novel prevention, diagnostic, and treatment strategies for an illness, or even suggestions for public health policy. The insights would be gained from predictions based on patterns among data. This means whole populations needn't be interviewed, nor even specifically sampled for a purpose. The data processing appears to do it alone. Nevertheless, in order to curate, populate, maintain, and operationalize such datasets

with sufficient quality, expert knowledge is required across a number of disciplines.

Data scientists, software and hardware engineers, and developers are needed to ensure quality systems and structures. Medical expertise is also required, able to identify health-relevant data and connections between that and other data. Likewise, to point out what is irrelevant, meaningless, obviously wrong, etc. Clinical expertise aside, there are also technical decisions that must be made about data which can bear upon their medical relevance. Indeed, how and in what respects data are accurate to the medical phenomena they represent is a question in need of careful scrutiny. As other contexts have shown, data isn't neutral with respect to its collection, storage, or mediation [12, 13].

Data-Utopianism

In an idealized data-utopian scenario, the AI approach to medicine would constitute what has been in another context termed a “screen and intervene” paradigm [14]. This is not “diagnosis” as it is known today, where interview, examination, and observation are essential elements. Biomarkers are of central importance in a more datafied approach. AI will look for these as patterns in data, on the basis that over time they have been established as indicators of illness. But biomarkers are not without problems in themselves, owing at least in part to questions over consistency and standardization [15]. The reduction to biomarkers represented as patterns in data, apt for automated detection, bypasses pertinent questions including some concerning what is being detected, and why that is being looked for in the first place (see section “Data Curation and Use” below).

It might be that the promises of AI prompt the datafication of medical investigation too quickly. If the data is approached in a spirit of exploration, and so too is the diagnostic field, then there appear to be overlapping explorations without specific guiding strategies. In human behavior generally, Tversky and Kahneman provide examples of decision-making under uncertainty wherein it is not necessarily helpful to gain more and more

information [16]. That is, when the future is uncertain, those facing a decision can be hampered in making optimal choices if they are presented with more to think about. Yet in this context of multiple uncertainties and explorations, the solution is taken to include amassing data. This is not without risk, especially in medicine where human health and wellbeing is at stake.

Unlike examples of human reasoning examined by Tversky and Kahneman, the AI approach doesn't proceed by gathering information and then coming up with decisions. In certain respects, AI centering on data turns traditional medical-scientific investigation on its head. Scientific research is often thought of in terms of heavily disciplined normal human reasoning: hypothesis-formation, investigation, then confirmation or refutation in the light of evidence gathered specifically for a carefully defined purpose. A typical way to characterize a clinical encounter might be along these lines: a patient presents with a complaint. The clinician carries out an examination, and makes observations. Based on these, and expert knowledge, a clinical diagnosis is made, following which, treatment or further examination is recommended. This stresses interpersonal skills, like empathy and including patients in the process overall [17]. The data-centered paradigm emphasizes a more thoroughly empirical drive, in which facts, in the shape of digitized data, are prioritized over testimony.

This approach instead amasses data from a variety of sources, and seeks hypotheses based on what emerges from the data. This is an exploratory approach that claims to need no guiding theory, instead seeking correlative information among data to prompt explanatory work. The models to explain patterns are "born from the data" [18]. Such approaches have seen some success in areas like predictive data analytics to forecast likely Parkinson's Disease development [19]. The data can also be harnessed to make narrow predictions about disease development for specific patients. Again, using Parkinson's as an example, whether a patient is likely to suffer falls or not can be taken from analysis of swathes of data far beyond the scale of comprehension of an individual clinician [20]. This kind of

predictive power concerning likely disease course is made available by data, and seems a dimension not open to the interpersonal clinical encounter. How to conceptualize the potential pros and any emerging cons of this approach, and how to weigh them against one another, remains a difficult endeavor. The following sections explore this difficulty in order to throw some light on what otherwise remains obscure.

Data Curation and Use

The aggregation of heterogeneous, large datasets, taken from myriad sources, makes the data in these applications complicated to deal with. This raises ethical questions about data ownership, privacy, consent, purpose, re-use, anonymity, and others, as well as the nature of dataset-making as a scientific, sociopolitical, and technological endeavor [21]. Once created, the analysis of these huge and varied datasets cannot be carried out by humans. Owing to the scale and complexity of data, algorithms and machine processing techniques more widely are required. Using techniques like Compressive Big Data Analytics (CBDA), algorithmic processing of data sets aims to be "model free" or "model agnostic" [22]. This kind of analysis can be taken to imply the objective nature of the data being collected, and of its subsequent analysis. But this objectivity is anything but secure.

One of the largest medical AI datasets at the moment is known as "ChestX-ray14." As Kulkarni et al. discuss [6] this is an interesting case that illustrates some of the issues raised here. This dataset was used in a study to train an AI-detection model called "CheXNet." In order to know what CheXNet should look for, a "ground truth" had to be established such that positive cases might be defined. This would represent what CheXNet was looking for. Ground truthing was carried out by text mining clinicians' radiology reports. This is the inclusion of medical expert opinion in the course of training the AI, mentioned above as necessary. But, "Intriguingly, CheXNet's performance mirrored human weaknesses in many respects; the algorithm had much

greater accuracy in detecting hiatal hernias, a radiographically distinctive diagnosis, compared to pulmonary infiltration, which is frequently ill-defined” [6].

This is taken to be the case because the notes from which the dataset is comprised, and on which the AI diagnostic model was trained, themselves are replete with uncertainty. Diagnoses aren’t always binary cases. One can speculate that the kinds of uncertainty present throughout clinical notes are exactly the sort of thing expertise is well honed to draw conclusions from, or otherwise use in informed medical decision-making. If the data is all there in the dataset, and machine learning techniques discover patterns in that set, then there is a ring of objectivity to it. The pattern is really there, somehow. But there are uncertainties in the data that are themselves hallmarks of medical expertise. Radiologists rightly record their uncertainties in their notes. What’s more, besides these kinds of responsible uncertainties, there are other potentially unconscious or historically derived biases in data. This is especially so where women and people of color are poorly represented in health data. Successful, data-driven skin cancer lesion detection algorithms, for instance, are effective mainly for light-skinned people [23]. The dataset here is incomplete following specific sociopolitical non-inclusivity. Issues such as these may be compounded if they become hidden behind the supposed objectivity of data processing.

Just as there are questions about dataset creation and its completeness, illustrated here with ChestX-ray14, others arise around algorithmic processing (see next section). How these questions and how the requirements of data in general relate to medical reasoning is in need of careful scrutiny. It is not clear that by deploying data in their investigations, clinicians will necessarily get closer to better answers simply owing to datafication.

There remain those who emphasize that the use of data ought to be handled judiciously, and used in tandem with expert clinical decision making [24]. Here, tools are deployed as decision support rather than considered as silver bullets. As with the scope for predicting falls in Parkinson’s

patients, data here can provide value. A clinician, faced with specific observations about their patient can draw upon arrays of information relating to similar observations. The singular case can thereby be related to a population-level sample. This could aid clinicians and patients alike in providing a rich backdrop to the otherwise one-to-one clinical encounter. It could further serve to minimize healthcare disparities by boosting clinician confidence and performance, not least through eliminating more burdensome dimensions of clinical work [25].

This would ideally translate into better patient outcomes, with world-class healthcare data available in even the most under resourced locations. While this promise sounds worthwhile, it does entail some changes to expected medical practice. The biostatistical reduction of the patient to correlations among data points not only methodologically detaches the person from their body and his/her sociopolitical context, but also the functioning of that specific body from clinical observation. This is replaced by an aggregating view of bodily function and deficit. The clinical encounter then takes on the function of harmonizing the deficient with the aggregate, mediated in depersonalized data processes. This represents a challenge to the traditionally interpersonal doctor-patient relationship. Whether and how this represents a problem, or a medical advance is an open question, though one addressed below.

AI and Medical Epistemology: Limits, Risks, and Biases

There are certain inevitabilities with regard to the limitations of a technology, including that it will serve more or less narrow purposes, and that it will be defined by those who imagine, determine, fund, and build it. In those senses then, “limits” and “biases” can be understood in terms of the parameters within which a technology is designed and developed. This includes the aims, ambitions, and outcomes, as well as the structure within which it comes to be, e.g., financial and political. A preference to develop one method and not another would count as a nontrivial bias on that

account, and the development of a technology that can do one thing but not another would count as one of its not unreasonable limitations (cf. [26]). Such limits and biases need not be either inherently negative or positive, nor are they necessarily problematic. Whether a limit or a bias matters would likely be linked to whether the technology “works” in fulfilling its aims and objectives, considered in context, and whether its benefits outweigh any harms.

This section turns to those limitations and biases that are necessarily negative, and in those respects the concern is with problematic limitations and biases that need not occur. These weaken the overall quality of the technology and its impact, and may cause a variety of harms, including physical, psychological, social, and environmental. The argument is that these kinds of limitations and biases are tied to a number of predictable, sometimes inevitable, negative consequences and therefore also risks. There are many negative biases that impact on AI, some of which are discussed here under two categories. The first concerns human biases and prejudices, which includes the kinds of biases contained and expressed in language as well as in interpretative endeavors including perception. The second concerns computational biases, and these are biases that are contained in programming and algorithms. This latter variety emerges to some extent from human biases, whether implicit or explicit, such as in the selection or categorization of data.

Human Biases and Prejudices: Language and Interpretation

Biases are unavoidable to some extent, and they can be either implicit, e.g., unacknowledged or even to some extent unknown, or explicit, e.g., stated, acknowledged, and to some extent known. Whether a bias is implicit or explicit, it can be demonstrated in both language and interpretation, i.e., words that are chosen, used, and understood, as well as in judgement and action, i.e., in decisions made and resulting behaviors. These biases may be intentional or unintentional, and they can result in a variety of consequences. For instance,

they can be seen in a tendency to connect certain illnesses to gender or culture, ethnicity, or socio-economic context. There are many examples of these biases, especially negative, in medical practice.

For instance, it has been suggested that there is a link between diagnosis of borderline personality disorder (BPD) and gender, including the possibility that it is overdiagnosed in young women [27], and elsewhere there is evidence to suggest that schizophrenia is overdiagnosed in young black men [28]. Meanwhile the expectation that someone who is overweight will necessarily suffer as a result of their weight has led to non-weight-related illnesses and diseases being misdiagnosed, missed, and even ignored [29]. Bias can also be tied to judgements made about a person based on their perceived class and socio-economic circumstances, including as these intersect with ethnicity [30]. Such problematic judgements are sometimes equated with understanding and knowledge. For instance, biases tied to perceptions of ability and disability intersect with gender with the outcome that disabled people describe being ignored or dismissed [31]. For example, people with visible disabilities report a lack of recognition regarding their complex situations. This includes where the focus centres only the disability, to the neglect of other topics, such as basic checks for blood pressure and cholesterol, and tests related specifically to prevention of disease. Patients report a lack of effective communication, particularly acute for those patients with severe disabilities, which sometimes results in excessive communication between physicians and caregivers to the neglect of the patient’s own perspectives [32].

How much credibility is given to a speaker, and how much credence to their testimony, has clear, direct, and serious consequences in medical contexts. A lack of care to the specificity of a patient’s concrete situation can lead to epistemic injustice, whereby negative biases and prejudices impact on the scope within which the speaker’s credibility is assessed and what then follows in terms of time given to their account, as well as outcomes and decisions (cf. [33, 34]). The kinds of biases in the examples noted above might be explicitly or

verbally expressed as prejudices, or they can be seen in the terminology that is used and the actions that follow. For instance, where the term “hysterical” subsumes and thereby also neglects a whole set of (typically women’s) physical and psychological health conditions [35]. Biases can also be enacted without being expressed, for instance, in the framing of a physician’s expectations and actions, which may nevertheless be recognized by patients, and which can frame their experience of the medical experience, as per the examples described above.

Recognition of some key principles can help to avoid such problematic biases in medical contexts. First, that empirical observation is never neutral, and as such empirical positions need to be viewed as at least partly normative [29]. Second, that the clinical gaze of individual physicians is itself normatively structured. As Foucault [36] argues, “The clinical gaze is not that of an intellectual eye that is able to perceive the unalterable purity of essences beneath phenomena. It is a gaze of the concrete sensibility, a gaze that travels from body to body, and whose trajectory is situated in the space of sensible manifestation.” This concrete sensibility is tied to contexts with specificity, cultural meaning, and, inevitably, more or less prejudice. Third, embodied humans have tacit bodily knowledge, which can be considered in terms of intercorporeal ways of knowing, i.e., as beings-in-the-world ([29] *ibid.*). It will suffice for now to highlight that the necessarily concrete nature of subjective experience both feeds into and is informed by biases. Even if are ways in which to recognize and mitigate these. The suggestion here is that these principles can also inform how AI is developed and used in medical contexts.

Computational Biases: Programming and Algorithms

Without careful attention, biases and prejudices can feed into both the data (see previous section) and the production of the algorithms on which AI relies. In such instances, a negative bias captured or reified in a technology like AI is always morally

problematic regardless of whether the technology is otherwise “successful” in its practical aims. An ethical approach to AI requires that people and ethics should not be sacrificed on the altar of “faster” or more efficient technologies, for instance benefiting some while harming others. It is worth recognizing that negative biases can, and often do, affect the overall success of a technology, but that conflating practical solutions with ethical solutions is not sufficient, and the ethical ought not to be subsumed within the practical. For instance, Google’s immediate response to the racism caused by their facial recognition algorithms – that identified some black people as gorillas – was to amend the recognition categories, i.e., by removing the identification of gorillas [37]. This technical workaround might solve a short-term practical problem, but it is very far from an ethical solution. The latter requires instead a broader recognition of the problems and their causes, as well as a wider range of solutions, some of which include greater representation both in terms of data and in the teams developing the AI [38].

To consider ethics first includes the recognition that human biases impact on computational biases, and that to some extent this may be unavoidable. It has been suggested that linguistic biases inevitably find their way into programming, and thereby into AI [39]. The argument suggests that biases are inherent to language, and since language is the framework within which an AI is developed and structured, it is also inevitable that bias will find its way into the programming. If this is accepted to be the case, then it becomes clear that vigilance for bias may not be enough. In medicine, this is especially problematic, and there are already many examples of where harm can occur if medical practitioners bring prejudices and biases to their practice, as discussed above. It is therefore essential to take this into account when planning, designing or using AI, if the risk of simply replicating and reifying those same biases is to be avoided in the development and use of these new tools. Where AI is used for automated decision-making and judgement, rather than as a tool for data management, the risk is even greater given the possibility that such systems will become embedded in medical

structures and future changes may be difficult or even impossible.

Once it is accepted that bias may be unavoidable, mitigation of negative biases in particular needs to be considered. For instance, it is clear that data scientists and engineers will not necessarily have expertise about the medical fields for which an AI is being developed. Evidence for this can be found in the language about medical conditions in papers that primarily describe the development of AI systems for use in those fields, e.g., the development of AI systems for the assessment of autism (cf. [40]). AI developers working on technologies for medical applications may demonstrate cursory understanding at best, and limited or flawed interpretations at worst. It's clear that the task of developing AI for fields outside of computing and engineering ought to be an interdisciplinary endeavor. Yet when expert knowledge is sought, this can lead to a replication of the biases of those whose theories are then privileged. As with the biases noted above, this may be inevitable to some extent. An expertly-informed AI system can serve to amplify the practice of the experts chosen to inform it. The views and preferences in theory and practice that are evident in an expert's perspective are implicit or explicit choices, judgements, and decisions. In selecting an expert to provide input to an AI system's development, elements of that specific perspective are being tacitly endorsed. Selecting expert advice is thereby, to some extent, omitting the theories and ideas that might animate the practice of an equally expert, but contrasting practitioner.

The above situation is not itself unusual, in so far as any clinician would also bring their preferences and biases to their individual practice. A key difference, however, concerns the scope and the reach that can be achieved by an embedded AI. In other words, what happens once a technology becomes ubiquitous? In the case of the individual physician, their prioritizing of certain theories, to the neglect of others, in their individual practice may only impact on the number of people who are patients in that practice, as well as whatever influence they may have on those they train or mentor. An AI that is developed with their

input would extend the reach of those biases and preferences, while also lending credence and authority to the selected theories. Autism is again a useful example here, especially as dominant theories about traits and characteristics have already led to exclusionary diagnostic practices [41]. Yet many of those problematically exclusionary practices are already being replicated in AI [40].

Mitigation requires a number of factors. First, that expertise is sought in the development of an AI, but also that such expertise is handled critically by an interdisciplinary team with a breadth of knowledge. This range ought to be sufficient for both the identification of problems of negative bias early in the design and implementation stages, as well as to ensure that sufficient attention is given to whether theories on which AI programming rely avoid exclusionary practices. Transparency in such processes is also essential so that medical professionals who will be end users of an AI enabled technology can themselves identify biases in the programming. Without these elements, the quality of an AI will remain vulnerable to the replication of unchecked biases, including those that are dominant but not unproblematic and those that have form for historical inequalities, whether in terms of overrepresentation or exclusion, among others. Opportunities to change the programming of potentially expensive AI may be few once it has already been embedded, so understanding of these necessary limitations and risks are essential if the tools are to be used critically and applied cautiously.

AI and Medical Ethics

The previous two sections explored how AI is transforming medical epistemology and bringing particular value to the field but also challenges and risks. This changing epistemological foundation with potential threats is also having a significant impact on medical ethics. The last section of this chapter is therefore dedicated to exploring how AI is affecting medical ethics.

Medicine is fundamentally a relational practice founded on the relationship between a patient and

a doctor. A key aspect of this relation is its constitutive asymmetry – the vulnerability of a patient who asks for the help of a health professional equipped with the competence to respond to the identified vulnerability. However, this asymmetry of abilities, or power, may lead to abuses on either side: (a) the patient might abuse the professional's service or (b) the professional might abuse his or her power over the patient. As the philosopher Worms puts it: "Care cannot exist without a relationship in which the weakness of one person requires assistance, which can turn into submission, and the capability of another allows for devotion, which can turn into power, or even abuse of power" ([42] ii). Because of this constitutive asymmetry between the patient and the doctor, medical practice has been, from very early on, regulated by ethical codes and guidelines. The earliest of these is the Hippocratic Oath and many others have followed, notably the principles of biomedical ethics by Beauchamp and Childress [43].

How is the deployment of AI in medicine modifying this fundamental aspect of this practice as a relation, given that it is one that is essentially asymmetric? This is a key question for medical ethics in the era of AI, one that philosophers, ethicists, healthcare professionals, patients, and society at large need to explore to identify potential challenges and risks and propose mitigating strategies. This section explores this question through two different aspects: (a) the patient-doctor relationship and (b) the medical profession in the era of digital capitalism.

The Patient-Doctor Relationship

The fundamental relationship between the patient and the doctor is affected in various ways by the introduction of AI to medicine. There is the risk that the clinical encounter – this fundamental moment of the therapeutic relationship – is set aside, replaced by a mass amount of data automatically collected and analyzed. The claimed "superhuman" accuracy and insight" [7] produced by AI risks replacing the value of this encounter. This is a potential threat for clinical

diagnosis as an outcome from a doctor-patient encounter. Indeed, what causes a person to seek medical help may not be the primary issue for which they need most help. In other words, the narrow approaches to diagnostic processes as conducted by AI systems may lead to missing potential for incidental findings. For instance, referred pain, physical manifestations of mental health issues, hoarding and other self-harm behaviors can be indicators of a brain disorder, a mental health condition, or of abuse, such as domestic violence. These kinds of potentialities could be straightforwardly considered in an interpersonal clinical encounter, but missed by an AI.

Additionally, with the entrance of new, highly complex technical tools in medical practice, the doctor might not fully understand the results of the analysis produced by the AI, and therefore be unable to properly explain to the patient and their family the full rationale behind the diagnosis and prescription. As Campolo and Crawford put it, "claims about 'superhuman' accuracy and insight" are "paired with the inability to fully explain how these results are produced" ([7] ibid). This is a key challenge of explainability brought about by AI, also presented as the issue of an AI system as a black box [44]. This issue is particularly critical in the context of medicine since human life and wellbeing are at stake.

The clinical encounter between the patient and doctor is also affected by the introduction of new actors, i.e., data scientists and engineers, and their technical systems, especially where these gain a central role. This inclusion affects the intimate relationship that characterizes the clinical encounter and brings challenges in terms of confidentiality and privacy [45, 46]. For instance, a study has shown that "a well-trained deep learning system is able to recover the patient identity from chest X-ray data" [47]. Even if CheXNet, discussed above, were to become an effective prediction model, such revelations could undermine its usefulness in terms of patient willingness to undergo scanning. This also leads to challenges related to trust and informed consent: two essential aspects of the medical relationship. In turn, this threat to the patient-doctor dialogue and relation of trust might make it difficult for the patient to make

sense of their own illness or issue. Indeed, the interpersonal dialogue with the doctor plays a major role in this self-understanding.

The Medical Profession in the Era of Digital Capitalism

The entrance of AI in medicine is also impacting the medical profession in various ways. Data science is increasingly gaining a central role in the field, at the expense of more traditional medical expertise, including its intuitive dimensions. As Matuchansky puts it: “One quickly learns from clinical practice that medicine is as much an art as a science or a technique” [48]. With the entrance of AI in medicine, this should not be forgotten or obscured behind claims of precision, scientism, and objectivity of AI systems.

Connected to this challenge to the profession, there is also the risk of deskilling of medical professionals [49]. If AI systems are seen as better placed to conduct various activities that were initially conducted by a doctor, chances are high that they will progressively replace humans (hence, also redirecting money from clinician’s wages to technology companies). For instance, use of deep learning techniques for chest radiography shows “the potential to exceed human performance” [47]. In turn, this leads to the risk of increased dependency on technology for key aspects of human existence (cf [49, 50]). Against this, many experts have called to ensure that AI remains an assistant to the healthcare professional, acting as a “support tool,” and not as a replacement to the human [51].

Additionally, this trend further pushes medicine in the direction of a technical discipline, taking attention away from the relational aspects of this practice. As the ethics of care has shown, the relational aspects of care already tends to be undervalued in the medical sector. These non-technical tasks are primarily undertaken by people from marginalized groups, primarily women and migrants, and those who are poorly paid. This can be contrasted with those highly specialized and technical areas of medicine such as surgery, which are highly respected and valued,

and rewarded accordingly [52]. The introduction of AI to the world of medical care further contributes to this over-valorization of the technical at the expense of relational aspects.

Exploring relations in medicine is not only restricted to interpersonal relationships (between patients, doctors, and other professionals). It also requires looking at relations within the broader sociopolitical landscape and the power asymmetries at this level. Here as well, the growing dependency on technology, and in particular on big technology companies needs to be investigated. The use of AI in the medical sector has brought about new actors, including Google, Apple, Facebook, and Amazon [53]. Because these companies are in possession of massive amounts of data and the ability to process them, they have found themselves well placed to enter the medical sector. For instance, Tamar Sharon talks about the “googlisation of healthcare” and its promises to “advance health research by providing the technological means for collecting, managing, and analysing the vast and heterogeneous types of data required for data-intensive personalised and precision medicine” [53]. This entrance of healthcare in “digital capitalism” is posing key questions in terms of privacy and confidentiality. Indeed, these big tech companies have a rather poor track record when it comes to the protection of personal data [50]. Medical data is enormously valuable and a particularly sensitive type of data that requires special protection. Access to such data only further increases the power of technology companies. Meanwhile, there is a general consensus that already powerful actors from insurance to recruitment companies should not have access to this data as it could lead to significant discrimination on the basis of health.

As the above has shown, the introduction of AI in medicine changes the power dynamics in the sector. With the digitalization of the field, data scientists, as well as big technology companies are gaining a central role in medicine. It is also pushing the sector toward “technicization” at the expense of relational aspects that are nonetheless central. Although AI in medicine brings great promises for the field in terms of efficiency and precision, it is nonetheless essential to pay

attention to these changing power dynamics as well as the technicization of the field to ensure that the interests and wellbeing of the patient remain the central consideration of medical practice. Additionally, it is essential to ensure that healthcare professionals, whether doctors or nurses, are well equipped for the digital transformation of their field and protected against abuses by industry actors. Finally, policy makers also have a role to play to ensure digital capitalism does not interfere with the interests of patients.

Conclusion and Recommendations

This chapter has investigated some of the promises for medicine in the algorithmic age and what these claims mean philosophically and ethically. After briefly exploring what AI promises to achieve in terms of medical advances, this chapter looked at how the introduction of AI in medicine deeply impacts medical epistemology. In particular, it explored the implications of a technology that claims to generate knowledge and diagnosis from data alone. Numerous questions raised by this epistemology, based on a form of “data-utopianism,” and how this interrogates the nature of objectivity in the medical field have been pointed out. This chapter has then highlighted the limitations, risks, and biases of AI and how these impact medicine. Finally, it looked into the ethical implications of the introduction of AI to medicine, especially in relation to the patient-doctor relationship and how the medical profession is evolving as it enters the era of digital capitalism.

This chapter has highlighted some key challenges and risks brought about by AI in medicine, but also some ways to mitigate these. To begin with, it is essential to refer to relevant ethical guidelines and frameworks. A number of these have been developed for AI over the last few years, such as the 2019 “Ethics guidelines for trustworthy AI” of the High-Level Expert Group on Artificial Intelligence set up by the European Commission [54]. The SIENNA project has also developed a set of ethical instruments to promote an ethical development, deployment and use of AI, including an ethics by design framework and an approach for research ethics [55]. It is also

important to carefully assess where AI can truly add value and where it does not (and might even be harmful). To these ends, ethical impact assessments as well as social science studies on the use of AI in medicine can help to better understand the impacts and consequences of AI for the medical sector, including in the short, medium, and long term [56].

Finally, it is essential for the healthcare community to develop their understanding of AI in technical terms. This would help to ensure appropriate and proportionate levels of trust, which is not too little, such that the AI system would become useless or poorly applied and adopted, nor too much, such that it is trusted and allowed to function independently of oversight and appropriate human intervention. It is also essential to raise awareness of the kinds of unique challenges that are brought by AI, including for doctors, for the healthcare community, and for society at large. Effective AI in medical contexts requires transparency so that people are aware of these tools in their work and their everyday lives, and so as to ensure a certain degree of oversight in these contexts. Finally, and as already mentioned above, AI should remain a tool that human beings can use if they deem it useful to achieve specific and identifiable objectives, but not as a replacement for human expertise and intervention, especially in medical domains where a person’s needs and their vulnerabilities may be greatest.

References

1. Coeckelbergh M. *AI ethics*. The MIT press essential knowledge series. Cambridge, MA: The MIT Press; 2020.
2. Mittelstadt B. AI ethics – too principled to fail? *SSRN Journal* [Internet]. 2019 [cited 2021 Mar 22]. <https://www.ssm.com/abstract=3391293>
3. Shaw JA, Sethi N, Block BL. Five things every clinician should know about AI ethics in intensive care. *Intensive Care Med*. 2021;47(2):157–9.
4. Schiff D, Borenstein J. How should clinicians communicate with patients about the roles of artificially intelligent team members? *AMA J Ethics*. 2019;21(2):E138–45.
5. Morley J, Machado CCV, Burr C, Cowls J, Joshi I, Taddeo M, et al. The ethics of AI in health care: a mapping review. *Soc Sci Med*. 2020;260:113172.

6. Kulkarni S, Seneviratne N, Baig MS, Khan AHA. Artificial intelligence in medicine: where are we now? *Acad Radiol.* 2020;27(1):62–70.
7. Campolo A, Crawford K. Enchanted determinism: power without responsibility in artificial intelligence. *Engag Sci Technol Soc.* 2020;6:1–19.
8. Anderson C. The end of theory: the data deluge makes the scientific method obsolete. *Wired.* 2008;16(07).
9. Good IJ. The philosophy of exploratory data analysis. *Philos Sci.* 1983;50(2):283–95.
10. Dinov ID. Volume and value of big healthcare data. *J Med Stat Inf.* 2016;4(1):3.
11. Butte AJ, Kohane IS. Unsupervised knowledge discovery in medical databases using relevance networks. In: *Proc AMIA Symp.* 1999;711–5.
12. van Dijck J. Datafication, dataism and dataveillance: big data between scientific paradigm and ideology. *Surveill Soc.* 2014;12(2):197–208.
13. danah b, Crawford K. Critical questions for big data. *Inf Commun Soc.* 2012;15(5):662–79.
14. Rose N. ‘Screen and intervene’: governing risky brains. *Hist Hum Sci.* 2010;23(1):79–105.
15. Poste G. Bring on the biomarkers. *Nature.* 2011;469(7329):156–7.
16. Tversky A, Kahneman D. Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychol Rev.* 1983;90(4):23.
17. Frankel RM, Stein T. Getting the most out of the clinical encounter: the four habits model. *Perm J.* 1999;3(3):79–88.
18. Kitchin R. Big data, new epistemologies and paradigm shifts. *Big Data Soc.* 2014;1(1):2053951714528481.
19. Dinov ID, Heavner B, Tang M, Glusman G, Chard K, Darcy M, et al. Predictive big data analytics: a study of Parkinson’s disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations. *PLoS One.* 2016;11(8):e0157077.
20. Gao C, Sun H, Wang T, Tang M, Bohnen NI, Müller MLTM, et al. Model-based and model-free machine learning techniques for diagnostic prediction and classification of clinical outcomes in Parkinson’s disease. *Sci Rep.* 2018;8(1):7129.
21. Mittelstadt BD, Floridi L. The ethics of big data: current and foreseeable issues in biomedical contexts. *Sci Eng Ethics.* 2016;22(2):303–41.
22. Marino S, Xu J, Zhao Y, Zhou N, Zhou Y, Dinov ID. Controlled feature selection and compressive big data analytics: applications to biomedical and health studies. *PLoS One.* 2018;13(8):e0202674.
23. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. *JAMA Dermatol.* 2018;154(11):1247–8.
24. Bezemer T, de Groot MCH, Blasie E, ten Berg MJ, Kappen TH, Bredenoord AL, et al. A human(e) factor in clinical decision support systems. *J Med Internet Res.* 2019;21(3):e11732.
25. Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care. *AMA J Ethics.* 2019;21(2):167–79.
26. Kuhn TS, Hacking I. The structure of scientific revolutions. 4th ed. Chicago/London: The University of Chicago Press; 2012. 217 p.
27. Becker D. Through the looking glass: women and borderline personality disorder [Internet]. 1st ed. Routledge; 2019 [cited 2021 Mar 23]. <https://www.taylorfrancis.com/books/9780429964206>
28. Metzl JM. The protest psychosis: how schizophrenia became a black disease [Internet]. Boston: Beacon Press; 2014 [cited 2021 Mar 23]. <http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=715745>
29. Murray S. Corporeal knowledges and deviant bodies: perceiving the fat body. *Soc Semiot.* 2007;17(3):361–73.
30. van Ryn M, Burke J. The effect of patient race and socio-economic status on physicians’ perceptions of patients. *Soc Sci Med.* 2000;50(6):813–28.
31. Olkin R, Hayward H, Abbene MS, VanHeel G. The experiences of microaggressions against women with visible and invisible disabilities. *J Soc Issues.* 2019;75(3):757–85.
32. Hamilton N, Olumolade O, Aittama M, Samoray O, Khan M, Wasserman JA, et al. Access barriers to healthcare for people living with disabilities. *J Public Health (Berl)* [Internet]. 2020 Oct 10 [cited 2021 Mar 23]. <http://link.springer.com/10.1007/s10389-020-01383-z>
33. Fricker M. Epistemic injustice: power and the ethics of knowing. Oxford/New York: Oxford University Press; 2007. 188 p.
34. Peled Y. Language barriers and epistemic injustice in healthcare settings. *Bioethics.* 2018;32(6):360–7.
35. Tasca C, Rapetti M, Carta MG, Fadda B. Women and hysteria in the history of mental health. *CPEMH.* 2012;8(1):110–9.
36. Foucault M. The birth of the clinic: an archaeology of medical perception. 1. publ., reprinted. London: Routledge; 2010. 266 p. (Routledge classics).
37. Vincent J. Google ‘fixed’ its racist algorithm by removing gorillas from its image-labeling tech [Internet]. The Verge. 2018 [cited 2021 Mar 24]. <https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai>
38. Garcia M. Racist in the machine: the disturbing implications of algorithmic bias. *World Policy J.* 2016;33(4):111–7.
39. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. *Science.* 2017;356(6334):183–6.
40. Erden YJ, Hummerstone H, Rainey S. Automating autism assessment: what AI can bring to the diagnostic process. *J Eval Clin Pract.* 2020;27:485. <https://doi.org/10.1111/jep.13527>.
41. Bargiela S, Steward R, Mandy W. The experiences of late-diagnosed women with autism spectrum conditions: an investigation of the female autism phenotype. *J Autism Dev Disord.* 2016;46(10):3281–94.
42. Worms F. The two concepts of care. *Life, medicine, and moral relations.* Esprit. 2006;1:141.
43. Beauchamp TL, Childress JF. Principles of biomedical ethics. Oxford: Oxford University Press; 2012.

44. Castelvecchi D. Can we open the black box of AI? *Nat News.* 2016;538(7623):20.
45. Char DS, Shah NH, Magnus D. Implementing machine learning in health care – addressing ethical challenges. *N Engl J Med.* 2018;378(11):981–3.
46. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med.* 2018;169(12):866–72.
47. Packhäuser K, Gündel S, Münster N, Syben C, Christlein V, Maier A. Is medical chest X-ray data anonymous? arXiv:210308562 [CS, EESS] [Internet]. 2021 Mar 15 [cited 2021 Mar 24]. <http://arxiv.org/abs/2103.08562>
48. Matuchansky C. Intelligence clinique et intelligence artificielle. Une question nuance med/sci. 2019;35: 797–803.
49. Susskind RE, Susskind D. The future of the professions: how technology will transform the work of human experts. Oxford University Press; 2015.
50. Powles J, Hodson H. Google DeepMind and healthcare in an age of algorithms. *Heal Technol.* 2017;7(4):351–67.
51. Coeckelbergh M. Health care, capabilities, and AI assistive technologies. *Ethical Theory Moral Pract.* 2010;13:181–90.
52. Molinier P. De la civilisation du travail à la société du care. *Vie sociale.* 2016;14(2):127–40.
53. Sharon T. When digital health meets digital capitalism, how many common goods are at stake? *Big Data & Society.* 2018.
54. Hleg A. High-level expert group on artificial intelligence: ethics guidelines for trustworthy AI. European Commission, 0904; 2019.
55. Resseguier A, Brey P, Dainow B, Drozdewska A, Santiago N, Wright D. D5.4: multi-stakeholder strategy and practical tools for ethical AI and robotics. SIENNA; 2021.
56. Resseguier A, Rodrigues R. Ethics as attention to context: recommendations for the ethics of artificial intelligence. Open Research Europe; 2021.



Reporting Standards and Quality Assessment Tools in Artificial Intelligence–Centered Healthcare Research

27

Viknesh Sounderajah, Pasha Normahani, Ravi Aggarwal,
Shruti Jayakumar, Sheraz R. Markar, Hutan Ashrafiyan, and
Ara Darzi

Contents

Introduction	386
The Case for AI-Specific Instruments	387
Specific AI Reporting Standards	389
SPIRIT-AI and CONSORT-AI	389
STARD-AI	389
TRIPOD-AI	392
Specific AI Quality Assessment Tools	392
QUADAS-AI	392
PROBAST-AI	393
Conclusion	393
References	393

Abstract

V. Sounderajah (✉) · R. Aggarwal · A. Darzi
Department of Surgery & Cancer, Imperial College London, London, UK

Institute of Global Health Innovation, Imperial College London, London, UK
e-mail: vs1108@imperial.ac.uk

P. Normahani · S. Jayakumar · S. R. Markar
Department of Surgery & Cancer, Imperial College London, London, UK

H. Ashrafiyan
Department of Surgery and Cancer, Imperial College London NHS Trust, London, UK

Institute of Global Health Innovation, Imperial College London, London, UK

Hamlyn Centre for Robotics and Artificial Intelligence,
Department of Surgery and Cancer, Imperial College London, London, UK

The practice of incomplete study reporting is rife within scientific literature. It hinders the adoption of technologies, introduces considerable “research waste,” and represents a significant moral hazard. In order to combat this issue, there has been a shift towards the use of reporting standards and quality assessment tools, a move that has been endorsed by major biomedical journals as well as other key stakeholders. These instruments help [1] to improve the quality and completeness of study reporting as well as [2] to aid researchers in their assessment of a study’s risk of bias and applicability. These instruments are carefully created through a multistep evidence generation process and are specific to individual study

designs or specialties. Recently, it has been noted that many of the existing instruments are poorly suited to aid the reporting and assessment of artificial intelligence (AI)-based studies on account of their niche study considerations. As such, there has been a concerted effort to produce AI-specific extensions to preexisting instruments, such as CONSORT, SPIRIT, STARD, TRIPOD, QUADAS, and PROBAST. This chapter expands upon why AI-specific amendments to these instruments are required in addition to highlighting their contents and proposed scope.

Keywords

Artificial intelligence · Reporting standards · Quality assessment tools · CONSORT-AI · SPIRIT-AI · STARD-AI · TRIPOD-AI · QUADAS-AI · PROBAST-AI

Introduction

Ensuring that treatment strategies are founded upon well-reported and unbiased scientific research is one of the cornerstones of good patient care. While completeness of reporting constitutes sound and ethical conduct among all scientific endeavors, it is of particular importance within health sciences, a field in which indiscretions may manifest in avoidable harm to patients. There are many notable examples of poor reporting, even within high profile randomized controlled trials, which have hindered the clinical translation of potentially useful research outcomes [1]. Such practice constitutes “research waste” [2] and poses a significant moral hazard.

As greater awareness regarding this issue has developed, experts have highlighted a number of key phases within a research study during which poor reporting may occur, as illustrated in Table 1.

While many hold the assumption that the responsibility of ensuring comprehensive research reporting falls primarily upon the editorial boards of journals, this, in fact, is a fallacy. Poor reporting is best viewed as a system failure across researchers, research funders, peer

Table 1 A table detailing examples of poor reporting across different study phases

Study phase	Examples of poor reporting
Study protocol	1) Ambiguous sample size calculation [3] 2) sparse administrative information regarding study conduct [4]
Main study abstract	1) Selective presentation of study results [5] 2) incongruity with message from the main study manuscript [6]
Main study methods	1) Ambiguous inclusion and exclusion criteria [4] 2) Inappropriate statistical analyses [7] 3) Inconsistencies in comparison to the associated study protocol [8]
Main study results	1) Incomplete reporting of study results [9] 2) Statistical errors [10] 3) Post hoc subgroup analyses [3] 4) Omission of harm reporting [11] 5) Misleading visual representation of data [12]
Main study discussion	1) Selective or inappropriate citation of other studies [13]

reviewers, and journal editorial boards, an issue, at least in part, analogous to the “Swiss cheese model” used in risk analysis across multiple industries [14]. As evidenced by the many thousands of dubious studies that have been published following peer review process [15], it is inappropriate and unrealistic to assume that a single layer of defense against poor reporting will suffice.

While the issue of poor reporting is a multi-faceted problem, one of the principal means in which the scientific community may mitigate the risk of this is through adherence to consensus-achieved reporting standards and quality assessment tools. Since their introduction in the mid-1990s [16], these instruments have garnered considerable traction in the scientific community and are now mandated by the majority of major biomedical journals [17]. Reporting standards and quality assessment tools may be differentiated as follows:

- Reporting standards provide structure and advice upon the information required in a primary health research manuscript. They typically consist of a checklist of minimally

essential items, which are often completed alongside accompanying explanatory documents and flow diagrams in order to facilitate its appropriate usage. Typically, reporting standards are specific to either the proposed study design or the clinical specialty under investigation. Moreover, each set of standards is created through a multistep evidence generation process. The aim of these processes, which often take years to complete, is to produce an internationally accepted set of reporting standards which aids the interpretation of the study's methodology, internal validity, external validity, and the overall "real world" utility. Given the needs across specific clinical specialties and study designs, there are now over 250 reporting standards available for consultation. In order to aid researchers in the appropriate identification, use, and creation of reporting standards, the Enhancing the QUAlity and Transparency Of health Research (EQUATOR) network [18] was established in 2008.

- (b) Quality assessment tools are instruments that are designed to specifically appraise one or more dimensions of quality within a research study. These tools are typically employed when conducting secondary research projects, such as systematic reviews and meta-analyses; study designs that are widely trusted to be the most reliable form of evidence within most taxonomies of empirical strength [19]. Systematic reviews should involve "a systematic approach to minimizing biases and random errors which is documented in a materials and methods section" [20]. In keeping with the aforementioned statement, a key component in conducting such secondary research studies is the appraisal of the risk of bias and applicability within shortlisted studies. Foregoing this crucial step can lead to the improper inclusion of studies which may, in turn, lead to unreliable and inaccurate estimates of treatment effects. Much like reporting standards, quality assessment tools are study design specific and are created following a multistep evidence generation process, typically led by experts within the field.

The Case for AI-Specific Instruments

As noted in other chapters in this book, AI-based technologies dominate medical headlines and are often touted as the panacea for a number of longstanding deficiencies across health systems globally [21]. Stakeholders from healthcare, government, computer science, and industry backgrounds are confident that AI can be positioned to tackle (1) the high rate of avoidable medical errors [22], (2) clinical workflow inefficiencies [23], and (3) the current level of unsustainable resource utilization associated with contemporary healthcare provision [24]. Despite these lofty ambitions, the integration of AI into everyday clinical practice within the health sector has been limited thus far. Issues that preclude its meaningful clinical translation across health systems include sparse regulatory guidance [25], wavering patient acceptability [26], and, pertinent to this chapter, the paucity of AI-specific instruments to assess the emerging scientific evidence.

As highlighted previously, reporting standards and quality assessment tools serve as critical tools which allow stakeholders from varying backgrounds to uniformly appraise the validity of research findings. While there are a wide number of specific tools available, none are specific to AI methodologies. As such, their ability to interrogate key reporting and quality considerations in this genre of studies is limited.

This is highlighted in the largest systematic review to be undertaken assessing the performance of AI models in healthcare, led by Liu et al. (2019) [27]. The review highlights that AI models have a high degree of diagnostic accuracy across a range of specialties, imaging modalities and clinical workflows. However, they also highlight the large degree of methodological and reporting variation that exists in the literature currently. Examples of this variance include:

1. Terminology: The term "validation" is used inconsistently in the literature. In some instances, the term "validation," when used in isolation, was used to describe the process of testing the final model, whereas other research groups used the term to describe the process of fine-tuning models as part of its development.

2. Dataset: The majority of AI studies are reliant upon retrospectively collected data, which are often labeled for purposes other than AI model development (e.g., more commonly for clinical radiology or histology reports). As such, the criteria for either the presence or absence of a pathology are poorly defined. In addition, there is often minimal rationale as to how images are chosen for inclusion within the dataset. There is often no justification for the size of datasets. Moreover, pertinent details regarding multi-vendor image sources are commonly missing. Lastly, there was often a lack of transparency regarding the split of data amongst the training and test sets, particularly when stating whether test sets were out of sample.
3. AI model: Very few studies provided enough detail that would allow for meaningful replication and independent validation of the models. Moreover, details which impact “real world” translation, such as availability of the model as well as whether the model is static or dynamic, are also poorly reported.
4. Study Methods: Studies often undertook their assessment of their models in isolation, which is poorly reflective of clinical practice. Moreover, even as part of the “*in silico*” assessment, many studies typically did not undertake validation of their model in an external test set, a poor practice that has been shown to lead to inflated measures of model performance. In addition, very few studies compared the performance of AI models against expert human clinicians.

Adding to this growing body of evidence is a meta-research assessment conducted by this authorship group. Through our own work, we have examined the use of existing quality assessment tools in secondary research papers reporting upon the diagnostic performance of AI systems. We note that formal quality appraisal and risk of bias assessment was not uniformly applied in AI-based diagnostic accuracy systematic reviews. Despite being considered a prerequisite among many journals, only 74% of studies shortlisted performed any form of quality assessment, with 56% of reviews opting for the QUADAS-2 tool, the most widely used tool within its class. Further to this, we are able to demonstrate that there is

either high or an unclear risk of bias in all of the quality domains assessed. We note the following issues, as framed by the QUADAS-2 domains of assessment:

1. Patient selection: The principal factors leading to a high risk of bias within the patient selection domain include poor patient sampling technique as well as the inappropriate exclusion of data at either a patient or feature level. AI-based diagnostics require rigorous datasets representative of real-world characteristics to produce reliable and generalizable diagnostic results. Therefore, inappropriate exclusion of participants and/or features can affect the interpretation of AI results in a more significant way compared to conventional analysis and contribute considerably to bias in patient selection. Excluding pathologies that may be similar in characteristics or have overlapping features to the diagnoses of interest (e.g., exclusion of patients with inflammatory bowel disease when determining diagnostic accuracy of endoscopic detection of polyps) may lead to overestimation of the algorithm’s diagnostic accuracy as well as low generalizability and clinical utility of the algorithm.
2. Index test: The term “index test” relates to the AI model that is being evaluated against the reference standard. AI models developed using overlapping datasets can overestimate diagnostic accuracy in comparison to using external validation data. Conversely, a lack of exposure to multiple manifestations of a pathology can result in “The Frame Problem” [29] whereby seemingly obvious diagnoses may be missed by the model simply due to inexperience.
3. Reference standard: Determination of an appropriate reference standard, which serves as the “ground truth” in model development, requires consideration of the best available evidence. This process may involve amalgamating clinical, radiological, and laboratory data. There are multiple instances in which AI models are compared to a human reference standard as the gold standard while alternative tests providing higher sensitivity and specificity are feasible. This is illustrated by Harris et al. (2019) [30] who deemed that 32 out of

33 shortlisted studies reporting upon the diagnostic accuracy of AI models to analyze chest x-rays for pulmonary tuberculosis were of high risk of bias due to the reference standard relying upon human interpretation of the chest x-ray as opposed to the use of sputum culture confirmation.

4. Flow and timing: The flow and timing domain considers the interval between the index test and reference standard, similarities in the reference standard evaluation among all patients and the inclusion of all patients in the final analysis. We highlight that the use of different reference standards between positive and negative cases can pose challenges in AI-based studies. This is illustrated in the case of cancer studies in which histological confirmation is often utilized as part of the reference standard for confirming malignancy; however, obtaining biopsies from clearly benign lesions poses ethical and practical challenges, thereby necessitating the use of alternative confirmatory tests [31]. Moreover, utilizing different reference standards, such as follow-up alone in comparison to histology, may also result in verification bias.

Given these aforementioned issues, there is a concerted global effort to address these flaws, promote more complete reporting, and produce quality assessment tools that are better placed to appraise emerging literature. A number of research teams are undertaking AI-specific extensions to existing reporting standards and quality assessment tools. In the following section, this chapter will highlight these AI-specific reporting standards and quality assessment tools that have either been developed or are in the process of development currently.

Specific AI Reporting Standards

SPIRIT-AI and CONSORT-AI

The SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) 2013 [32] document serves as guidance for authors who wish to undertake either randomized or nonrandomized trials. It is an essential document that enables readers

to undertake three key tasks: (1) understand the background, methods, target population, statistical analyses, and ethico-legal considerations associated with the proposed study, (2) replicate aspects of the trial methodology, if desired, and (3) appraise the study's scientific rigor.

It is essential for study protocols to be complete in their reporting as they often serve as critical pieces of evidence when used by external reviewers, such as funding bodies, regulatory authorities, research ethics committees, and peer-reviewers on behalf of scientific journals. However, despite its comprehensiveness, SPIRIT 2013 does concede that certain methodologies and specialties would require a specific extension.

The CONSORT (Consolidated Standards of Reporting Trials) statement provides guidance for authors who are reporting the findings of a randomized controlled trial (RCT). It is well accepted that RCT designs are considered to be the gold standard in providing evidence upon the safety and efficacy of an intervention. Stakeholders often rely upon the results of RCTs in order to inform clinical standards and health policy.

CONSORT was first introduced in 1996 [16] and was last updated in 2010 [33]. Like the other statements in this section, the CONSORT study authors have conceded that certain methodologies and specialties would require a specific extension.

Both SPIRIT-AI [34] and CONSORT-AI [35] initiatives were conducted by the same multi-stakeholder research group in order to ensure harmonization between statements. The AI-specific extensions and elaborations to both statements are listed in Table 2.

STARD-AI

The STARD (Standards for Reporting of Diagnostic Accuracy Studies) 2015 statement [36] remains the most widely accepted set of reporting standards for diagnostic accuracy studies. However, STARD was not designed to address the issues and challenges raised by AI-driven modalities. This is an increasingly pressing issue due to the fact that the majority of AI interventions that are close to translation are predominantly in the field of medical diagnostics. Recent market

Table 2 A table detailing the corresponding AI-specific additions to SPIRIT-AI and CONSORT-AI

SPIRIT-AI Item	CONSORT-AI Item	Explanation
Item 1 (i) elaboration: Indicate that the intervention involves artificial intelligence/machine learning and specify the type of model	CONSORT-AI 1a,b (i) elaboration: Indicate that the intervention involves artificial intelligence/machine learning in the title and/or abstract and specify the type of model	<i>This facilitates indexing and searching. The use of the broader terms of “artificial intelligence” and “machine learning” within the title allow for easy interpretability by a wide audience</i>
Item 1 (ii) elaboration: State the intended use of the AI intervention	CONSORT-AI 1a,b (ii) elaboration: State the intended use of the AI intervention within the trial in the title and/or abstract	<i>This allows readers to understand the intended use of the AI intervention in an unambiguous manner</i>
Item 6a (i) extension: Explain the intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users (such as healthcare professionals, patients, public)	CONSORT-AI 2a (i) extension: Explain the intended use for the AI intervention in the context of the clinical pathway, including its purpose and its intended users (e.g., healthcare professionals, patients, public)	<i>A detailed description of the clinical workflow and the role of the AI intervention within this system aids readers in understanding the importance of the study</i>
Item 6a (ii) extension: Describe any preexisting evidence for the AI intervention	N/A	<i>Details regarding the development and subsequent validation of the intervention provide useful background and should be provided</i>
Item 9 extension: Describe the onsite and offsite requirements needed to integrate the AI intervention into the trial setting	CONSORT-AI 4b extension: Describe how the AI intervention was integrated into the trial setting, including any onsite or offsite requirements	<i>AI interventions are heavily dependent upon specific operational environments. As such, it is imperative that these are explicitly specified. This includes the presence of vendor-specific devices, hardware requirements, and software requirements</i>
Item 10 (i) elaboration: State the inclusion and exclusion criteria at the level of participants	CONSORT-AI 4a (i) elaboration: State the inclusion and exclusion criteria at the level of participants	<i>It is important to differentiate between the inclusion/exclusion criteria at both a participant level and a data input level</i>
Item 10 (ii) extension: State the inclusion and exclusion criteria at the level of the input data	CONSORT-AI 4a (ii) extension: State the inclusion and exclusion criteria at the level of the input data	<i>There should be clearly stated minimum input data entry requirements, which can be related to image resolution or data format</i>
Item 11a (i) extension: State which version of the AI algorithm will be used	CONSORT-AI 5 (i) extension: State which version of the AI algorithm was used	<i>AI interventions should state the version number associated with the study. If there are alterations with respect to the agreed upon version number, this should be justified. When available, regulatory marking references should be provided</i>
Item 11a (ii) extension: Specify the procedure for acquiring and selecting the input data for the AI intervention	CONSORT-AI 5 (ii) extension: Describe how the input data were acquired and selected for the AI intervention	<i>The quality of the AI intervention output is heavily dependent upon the quality and nature of the input data that is selected for the study. As such, clear procedures regarding acquisition, selection, and preprocessing should be provided in order to allow input standardization across study site.</i>

(continued)

Table 2 (continued)

SPIRIT-AI Item	CONSORT-AI Item	Explanation
Item 11a (iii) extension: Specify the procedure for assessing and handling poor quality or unavailable input data	CONSORT-AI 5 (iii) extension: Describe how poor-quality or unavailable input data were assessed and handled	<i>As previously noted, poor quality data can severely impact upon the performance of the AI intervention. As such, minimum standards for input data are required. There should be transparency regarding how poor data is handled</i>
Item 11a (iv) extension: Specify whether there is human-AI interaction in the handling of the input data, and what level of expertise is required for users	CONSORT-AI 5 (iv) extension: Specify whether there was human-AI interaction in the handling of the input data, and what level of expertise was required of users	<i>This relates to how end users interact with the AI intervention and whether specific user training is required in order to attain maximal benefit from the intervention</i>
Item 11a (v) extension: Specify the output of the AI intervention	CONSORT-AI 5 (v) extension: Specify the output of the AI intervention	<i>The output, which can vary between a diagnostic classification, a recommended action, and a prognostic probability, needs to be clearly specified</i>
Item 11a (vi) extension: Explain the procedure for how the AI intervention's outputs will contribute to decision-making or other elements of clinical practice	CONSORT-AI 5 (vi) extension: Explain how the AI intervention's outputs contributed to decision-making or other elements of clinical practice	<i>This extends from the premise set forth in item 11a (iv) whereby there needs to be a clear process in place for the human-AI interaction when interpreting the intervention's outputs. There needs to be clear criteria as to who (experience and skill level) interprets the output as well as what the output is permitted to safely suggest (e.g., diagnostic probability expressed as a percentage vs. binary outcome values)</i>
Item 22 extension: Specify any plans to identify and analyze performance errors. If there are no plans for this, explain why not	CONSORT-AI 19 extension: Describe results of any analysis of performance errors and how errors were identified, where applicable. If no such analysis was planned or done, explain why not	<i>Failure case analysis is of vital performance as these nascent technologies become increasingly incorporated into clinical practice. This is a vital step in minimizing potential harm to patients. When appropriate, risk mitigation strategies should also be provided</i>
Item 29 extension: State whether and how the AI intervention and/or its code can be accessed, including any restrictions to access or re-use	CONSORT-AI 25 extension: State whether and how the AI intervention and/or its code can be accessed, including any restrictions to access or re-use	<i>This is an important part of maintaining transparency in study conduct</i>

analysis performed by Benjamins et al. (2020) [37] has demonstrated that over 90% of AI products that have been approved by the United States Food & Drug Administration are related to the field of diagnostics. In the current paradigm, diagnostic investigations require timely interpretation from an expert clinician in order to generate a diagnosis and to direct subsequent episodes of

care. The recurring issue with this system is that diagnostic services are inundated with large volumes of work, which can often supersede workforce capacity. In order to address this, diagnostic AI algorithms, with notable examples targeted toward breast cancer [38] and lung cancer [39], have positioned themselves as medical devices that may achieve diagnostic accuracy comparable

to that of an expert clinician while concurrently alleviating health-resource use. However, although this paradigm shift may seem imminent, it is crucial to note that much of the evidence supporting diagnostic algorithms has been disseminated in the absence of AI-specific reporting standards.

In order to tackle these problems, the STARD-AI Project Team and Steering Committee are preparing an AI-specific extension to the STARD statement (STARD-AI) [40] that aims to focus upon the specific reporting of AI diagnostic accuracy studies. This work is complementary to the other novel AI extensions noted in this section. This process is being developed in close collaboration with key stakeholders consisting of clinicians, computer scientists, journal editors, researchers, trialists, industry leaders, regulators, funders, policy makers, and patient groups. STARD-AI is anticipated in Q1 of 2022.

TRIPOD-AI

The TRIPOD (Transparent Reporting of a Multi-variable Prediction Model for Individual Prognosis or Diagnosis) statement was published in 2015 [41] and provides guidance on the key items to report when describing studies developing, evaluating, or updating clinical prediction models.

Although TRIPOD has been well adopted in the reporting of regression centered prediction model studies, its uptake within AI studies has been comparatively stunted thus far. This is a concerning feature given that one of the most prominent areas of AI healthcare research centers around the use of prognostic models. These models offer the potential of analyzing large and complex patient datasets in order to create precise and personalized outputs pertaining to patient prognosis [42]. This, in turn, can be leveraged in preventative clinical strategies which can lead to improved patient outcomes as well as more efficient resource utilization at a health system level.

In order to address these issues, TRIPOD-AI [43] is being developed in order to focus on the reporting of machine learning prediction algorithms. Like STARD-AI, TRIPOD-AI is being

developed in close collaboration with key stakeholders consisting of clinicians, computer scientists, journal editors, researchers, trialists, industry leaders, regulators, funders, policy makers, and patient groups.

Specific AI Quality Assessment Tools

QUADAS-AI

The most commonly used tool for the methodological assessment of diagnostic accuracy centered secondary research studies remains the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) tool [44]. QUADAS was developed in 2003 [45] and updated, as QUADAS-2, in 2011. The original tool comprised of 14 items on patient selection and spectrum, reference standard, presence of various biases, test execution, study withdrawals, and indeterminate results. The updated version was streamlined to categorize the questions into four key domains: (1) patient selection, (2) index test, (3) reference standard, and (4) flow and timing, with each domain assessed for the risk of bias as well as potential external validity. Although systematic reviews serve an important role in summarizing evidence in a precise fashion that aims to minimize bias, 96% of those related to AI diagnostic accuracy have been conducted in the absence of an AI-specific methodological quality assessment tool. The absence of a robust quality assessment tool for AI in this field not only hinders efficient quality appraisal at an evidence synthesis phase but has considerable downstream effects as key stakeholders; such policy makers, regulatory officials, technologists, and healthcare professionals are unable to effectively evaluate the translational potential of these nascent technologies.

In order to tackle this issue, an AI-specific extension to QUADAS-2 (QUADAS-AI [46]) is in production, which aims to provide researchers and policy makers with a framework to appraise methodological quality in systematic reviews evaluating the diagnostic accuracy of AI models. This work will be complementary to the ongoing STARD-AI initiative [40], PRISMA-DTA [47],

and QUADAS-3 and QUAPAS [48]. QUADAS-AI is being coordinated by a global Project Team and Steering Committee consisting of clinician scientists, computer scientists, epidemiologists, statisticians, journal editors, EQUATOR Network representatives, regulatory leaders, epidemiologists, statisticians, industry leaders, funders, health policy makers, legal experts, and bioethicists.

PROBAST-AI

PROBAST (Prediction model Risk Of Bias Assessment Tool) was published in 2019 [49] in order to help stakeholders critically appraise the study design, conduct and analysis associated with prediction model studies. PROBAST consists of four domains (participants, predictors, outcome, and analysis) and 20 signaling questions to facilitate risk of bias assessment. PROBAST-AI is being developed in parallel to the ongoing TRIPOD-AI initiative and will serve as a tool to assess risk of bias in machine learning-based prediction model studies.

Conclusion

There is increasing recognition that AI studies require bespoke reporting standards and quality assessment tools in order to aid completeness of reporting and minimization of bias. CONSORT-AI, SPIRIT-AI, and the forthcoming STARD-AI, TRIPOD-AI, QUADAS-AI, and PROBAST-AI will all serve first as consensus-derived iteration of global standards in this space.

References

- Casas J-P, Kwong J, Ebrahim S. Telemonitoring for chronic heart failure: not ready for prime time. In: Cochrane database of systematic reviews [Internet]. Wiley; 2010 [cited 2021 Mar 15]. Available from: <https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.ED000008/full>
- Glasziou P, Chalmers I. Research waste is still a scandal- A n essay by Paul Glasziou and Iain Chalmers. BMJ [Internet]. 2018 [cited 2021 Mar 15];363. Available from: <https://www.bmj.com/content/363/bmj.k4645>
- Chan AW, Hróbjartsson A, Jørgensen KJ, Gøtzsche PC, Altman DG. Discrepancies in sample size calculations and data analyses reported in randomised trials: comparison of publications with protocols. BMJ [Internet]. 2008 [cited 2021 Mar 15];337(7683):1404–7. Available from: <http://www.bmjjournals.org>/
- Glasziou P, Meats E, Heneghan C, Shepperd S. What is missing from descriptions of treatment in trials and reviews? [Internet]. BMJ. BMJ Publishing Group; 2008 [cited 2021 Mar 15];336:1472. Available from: <https://www.bmjjournals.org/content/336/7659/1472>
- Pitkin RM, Branagan MA, Burmeister LF. Accuracy of data in abstracts of published research articles. J Am Med Assoc [Internet]. 1999 [cited 2021 Mar 15];281(12):1110–1. Available from: [https://jamanetwork.com/](https://jamanetwork.com)
- Estrada CA, Bloch RM, Antonacci D, Basnight LL, Patel SR, Patel SC, et al. Reporting and concordance of methodologic criteria between abstracts and articles in diagnostic test studies. J Gen Intern Med [Internet]. 2000 [cited 2021 Mar 15];15(3):183–7. Available from: [/pmc/articles/PMC1495348/](https://pmc.ncbi.nlm.nih.gov/pmc/articles/PMC1495348/)
- Vesterinen H V, Egan K, Deister A, Schlattmann P, MacLeod MR, Dirnagl U. Systematic survey of the design, statistical analysis, and reporting of studies published in the 2008 volume of the Journal of Cerebral Blood Flow and Metabolism. J Cereb Blood Flow Metab [Internet]. 2011 [cited 2021 Mar 15];31(4):1064–72. Available from: <http://jcbfm.sagepub.com/doi/10.1038/jcbfm.2010.217>
- Dwan K, Altman DG, Blundell M, Gamble CL, Williamson PR. Comparison of protocols and registry entries to published reports for randomised controlled trials. In: Cochrane database of systematic reviews. Wiley; 2010.
- Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles [Internet]. JAMA. 2004 [cited 2021 Mar 15];291:2457–65. Available from: <https://pubmed.ncbi.nlm.nih.gov/15161896/>
- Ly WK, Strasak AM, Zaman Q, Pfeiffer KP, Göbel G, Ulmer H. Statistical errors in medical research-a review of common pitfalls [Internet]. Swiss Medical Weekly. 2007;37:0304. EMH Media; 2007 [cited 2021 Mar 15]. Available from: <https://smw.ch/article/doi/smw.2007.11587>
- Chowers MY, Gottesman BS, Leibovici L, Pielmeier U, Andreassen S, Paul M. Reporting of adverse events in randomized controlled trials of highly active antiretroviral therapy: systematic review [Internet]. J Antimicrob Chemother. 2009 [cited 2021 Mar 15];64:239–50. Available from: <https://pubmed.ncbi.nlm.nih.gov/19477890/>
- Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz LM, Woloshin S. Helping Doctors and patients make sense of health statistics. 2008.
- Jannet AS, Agoritsas T, Gayet-Ageron A, Perneger TV. Citation bias favoring statistically significant studies was present in medical research. J Clin Epidemiol

- [Internet]. 2013 [cited 2021 Mar 15];66(3):296–301. Available from: <https://pubmed.ncbi.nlm.nih.gov/23347853/>
14. Reason J. The contribution of latent human failures to the breakdown of complex systems. *Philos Trans R Soc Lond B Biol Sci* [Internet]. 1990 [cited 2021 Mar 15]; 327(1241):475–84. Available from: <https://pubmed.ncbi.nlm.nih.gov/1970893/>
 15. Altman DG. Poor-quality medical research: what can journals do? [Internet]. *JAMA*. 2002 [cited 2021 Mar 15];287:2765–7. Available from: <https://pubmed.ncbi.nlm.nih.gov/12038906/>
 16. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *J Am Med Assoc* [Internet]. 1996 [cited 2021 Mar 15];276(8):637–9. Available from: <https://pubmed.ncbi.nlm.nih.gov/8773637/>
 17. Article types and preparation | The BMJ [Internet]. [cited 2021 Mar 15]. Available from: <https://www.bmjjournals.org/about-bmjj/resources-authors/article-types>
 18. The EQUATOR Network | Enhancing the QUAlity and Transparency Of Health Research [Internet]. [cited 2020 Sep 26]. Available from: <https://www.equator-network.org/>
 19. OCEBM Levels of Evidence Working Group. The Oxford 2011 Levels of Evidence. Vol. 1, Oxford Centre for Evidence-Based Medicine. 2011. p 5653.
 20. Chalmers I, Altman D. Systematic reviews. London: BMJ Publishing Group Ltd. 1995.
 21. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future [Internet]. *Stroke Vasc Neurol*. BMJ Publishing Group. 2017 [cited 2021 Jan 17];2:230–43. Available from: <http://svn.bmjjournals.org/>
 22. Choudhury A, Asan O. Role of artificial intelligence in patient safety outcomes: systematic literature review [Internet]. *JMIR Med Inform*. JMIR Publications Inc.; 2020 [cited 2021 Jan 17];8. Available from: [/pmc/articles/PMC7414411/?report=abstract](https://pmc.ncbi.nlm.nih.gov/PMC7414411/?report=abstract)
 23. How AI Can Help Reduce \$200B in Annual Waste [Internet]. [cited 2021 Jan 17]. Available from: <https://www.optum.com/business/resources/library/artificial-intelligence-reduces-waste-health-care-costs.html>
 24. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. Nature Publishing Group. 2019;25:44–56.
 25. US Food and Drug Administration (FDA). Software as a Medical Device (SaMD) Action Plan [Internet]. 2021 [cited 2021 Mar 2]. Available from: www.fda.gov
 26. McCradden MD, Baba A, Saha A, Ahmad S, Boparai K, Fadaiefard P, et al. Ethical concerns around use of artificial intelligence in health care research from the perspective of patients with meningioma, caregivers and health care providers: a qualitative study. *C Open* [Internet]. 2020 [cited 2021 Mar 15];8(1):E90–5. Available from: [/pmc/articles/PMC7028163/](https://pmc.ncbi.nlm.nih.gov/PMC7028163/)
 27. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Heal*. 2019;1(6):e271–97.
 28. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *arXiv* [Internet]. 2019 [cited 2021 Mar 15]; Available from: <http://arxiv.org/abs/1908.09635>
 29. The Frame Problem (Stanford Encyclopedia of Philosophy) [Internet]. [cited 2021 Mar 15]. Available from: <https://plato.stanford.edu/entries/frame-problem/>
 30. Harris M, Qi A, Jeagal L, Torabi N, Menzies D, Korobitsyn A, et al. A systematic review of the diagnostic accuracy of artificial intelligence-based computer programs to analyze chest x-rays for pulmonary tuberculosis. *PLoS One* [Internet]. 2019 [cited 2021 Mar 15];14(9). Available from: <https://pubmed.ncbi.nlm.nih.gov/31479448/>
 31. Marka A, Carter JB, Toto E, Hassanpour S. Automated detection of nonmelanoma skin cancer using digital images: a systematic review. *BMC Med Imaging* [Internet]. 2019 [cited 2021 Mar 15];19(1):21. Available from: <https://bmcmedimaging.biomedcentral.com/articles/10.1186/s12880-019-0307-7>
 32. Chan A-W, Tetzlaff JM, Altman DG, Laupacis A, Götzsche PC, Krleža-Jerić K, et al. SPIRIT 2013 statement: defining standard protocol items for clinical trials. *Ann Intern Med* [Internet]. 2013 [cited 2021 Mar 15];158(3):200. Available from: <http://annals.org/article.aspx?doi=10.7326/0003-4819-158-3-201302050-00583>
 33. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ* [Internet]. 2010 [cited 2021 Mar 15];340(7748):698–702. Available from: <https://www.bmjjournals.org/content/340/bmj.c332>
 34. Rivera SC, Liu X, Chan A-W, Denniston AK, Calvert MJ. Consensus statement Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension The SPIRIT-AI and CONSORT-AI Working Group*, SPIRIT-AI and CONSORT-AI Steering Group and SPIRIT-AI and CONSORT-AI Consensus Group. *Nat Med* [Internet]. 2020 [cited 2020 Sep 26];26(9):1351–63. Available from: <https://doi.org/10.1038/s41591-020-1037-7>
 35. Liu X, Rivera SC. Consensus statement Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension 6,13 and The SPIRIT-AI and CONSORT-AI Working Group*. *Nat Med* 2020 269 [Internet]. 2020 [cited 2020 Sep 26];26(9):1364–74. Available from: <https://doi.org/10.1038/s41591-020-1034-x>
 36. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ*. 2015;28:351.

37. Benjamins S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *npj Digit Med* [Internet]. 2020 [cited 2020 Sep 26];3(1):118. Available from: <http://www.nature.com/articles/s41746-020-00324-0>
38. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafiyan H, et al. International evaluation of an AI system for breast cancer screening. *Nature* [Internet]. 2020;577(7788):89–94. <https://doi.org/10.1038/s41586-019-1799-6>.
39. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med* [Internet]. [cited 2020 Jul 2]; <https://doi.org/10.1038/s41591-019-0447-x>.
40. Sounderajah V, Ashrafiyan H, Aggarwal R, De Fauw J, Denniston AK, Greaves F, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: the STARD-AI Steering Group [Internet]. *Nat Med*. Nature Research; 2020 [cited 2020 Sep 26];26:807–8. <https://doi.org/10.1038/s41591-020-0941-1>.
41. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med* [Internet]. 2015 [cited 2021 Mar 15];13(1):1. Available from: <http://www.biomedcentral.com/1741-7015/13/1>
42. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. *npj Digit Med* [Internet]. 2018 [cited 2018 Jun 19];1. Available from: <https://www.nature.com/articles/s41746-018-0029-1.pdf>
43. Collins G, Moons K. Reporting of artificial intelligence prediction models. *Lancet*. 2019;393:1577–9.
44. Whiting PF. QUADAS-2: a Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann Intern Med* [Internet]. 2011 [cited 2021 Jan 17];155(8):529. Available from: <http://annals.org/article.aspx?doi=10.7326/0003-4819-155-8-201110180-00009>
45. Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews [Internet]. *BMC Med Res Methodol*. BioMed Central Ltd.; 2003 [cited 2021 Mar 2];3:1–13. Available from: <http://bmcmethodol.biomedcentral.com/articles/10.1186/1471-2288-3-25>
46. Sounderajah V, Ashrafiyan H, Deeks J, Whiting P, Bossuyt P, Collins G, et al. QUADAS-AI: a revised tool for the quality assessment of artificial intelligence centred diagnostic accuracy studies. 2021 [cited 2021 Mar 15]; Available from: <https://osf.io/fcpjt/>
47. McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, Clifford T, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies the PRISMA-DTA statement. *JAMA* [Internet]. 2018 [cited 2021 Mar 15];319(4):388–96. Available from: <https://pubmed.ncbi.nlm.nih.gov/29362800/>
48. Quality Assessment of Prognostic Accuracy Studies (QUAPAS): an extension of the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool for systematic reviews of prognostic test accuracy studies | Colloquium Abstracts [Internet]. [cited 2021 Mar 15]. Available from: <https://abstracts.cochrane.org/2019-santiago/quality-assessment-prognostic-accuracy-studies-quapas-extension-quality-assessment>
49. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* [Internet]. 2019 [cited 2021 Mar 2];170(1):51. Available from: <http://annals.org/article.aspx?doi=10.7326/M18-1376>



AIM and Gender Aspects

28

Didem Stark and Kerstin Ritter

Contents

Introduction	398
Sex and Gender Differences in Medicine	398
Sex and Gender Bias in Machine Learning Models	399
Role of Sex and Gender in Machine Learning Models for Medicine	400
Current Issues	401
Outlook and Potential Solutions	401
References	404

Abstract

As the use of machine learning (ML), a sub-field of artificial intelligence (AI), is becoming popular in healthcare, understanding the role of demographic traits such as sex, gender, age,

race, and socioeconomic status on model performance becomes crucial. In this chapter, we first give an introduction into the concepts of gender and sex and show how medical research has been biased toward these concepts. Since ML models are mostly based on historical medical datasets, existing bias might be picked up and distort ML results in favor of one group over another. After explaining the methods for finding bias in ML models, we present possible solutions for debiasing ML models including the collection of more balanced datasets, better documentation regarding underlying data as well as better documentation of the ML model itself, and the use of fairness metrics during model training. Finally, we discuss the current challenges and highlight the potential of deep neural networks for helping in mitigating bias.

D. Stark (✉)

Charité – Universitätsmedizin Berlin, Humboldt-Universität zu Berlin, Berlin, Germany

Department of Psychiatry and Psychotherapy, Berlin Institute of Health, Bernstein Center for Computational Neuroscience, Berlin, Germany
e-mail: didem.stark@bccn-berlin.de

K. Ritter

Department of Psychiatry and Psychotherapy, Charité – Universitätsmedizin Berlin, Berlin, Germany
Humboldt-Universität zu Berlin, Berlin, Germany

Department of Psychiatry and Psychotherapy, Berlin Institute of Health, Bernstein Center for Computational Neuroscience, Berlin, Germany

Keywords

Machine learning · Deep learning · Artificial intelligence · Gender bias · Sex bias · Gender in medicine · Sex in medicine

Introduction

Currently, ML models are expected to become a transformative tool in healthcare, bringing individualized medicine for everyone. As these ML models are used more and more extensively in everyday life, it becomes vital to evaluate the outcome performance disaggregated on various demographic traits and make sure that the ML models are fair and not discriminative across those traits. These demographic traits include but are not limited to sex, gender, age and race, as well as socioeconomic status. Here, we will focus on gender and sex traits in ML models used in medicine.

According to the World Health Organisation (WHO), sex can be defined as different biological characteristics of females and males, whereas gender can be defined as societal characteristics of those as well as individuals across nonbinary gender scales [1]. In addition to sex differences, gender also affects access to healthcare, and especially gender minorities have been discriminated against based on societal judgments [2, 3].

Although there are clear sex differences between males and females, for long years, sex differences were mostly ignored in medicine. Male physiology was accepted as the default, and most medical research was done using the male body as standard [4]. In the last decades, sex- and gender-specific approaches to medicine are gaining popularity, and more research is being done to understand sex- and gender-related differences in metabolism, physiology, and illnesses. Additionally, clinical trials are planned in a way such that they include both males and females [5]. Yet, sex and gender bias in medicine persists and is still a big problem. While some diseases are sex specific, such as various types of cancers affecting reproductive systems, others have different prevalence among males and females. For

instance, most autoimmune disorders such as multiple sclerosis (MS) are more common in females than males, whereas type 1 diabetes affects more males than females [6]. Additionally, some diseases such as cardiovascular diseases that affect both males and females manifest differently for each sex [7, 8]. However, as many clinical trials have been predominantly performed on male participants until recently, different symptoms in females could be overlooked and might result in females being underdiagnosed.

The gender and sex bias in medicine is currently present in the datasets that are used to train ML models for healthcare. The historical bias is not only reflected in the ML models but can even be amplified [9]. Recently, various methods for detecting and mitigating bias in ML models have been developed [10], and while they might not be specific to ML models used in healthcare, some of them have successfully been applied to medical data.

In the following parts, sex and gender differences in medicine, sex and gender bias in ML systems, and their interaction will be reviewed. Finally, current challenges and possible solutions will be discussed.

Sex and Gender Differences in Medicine

Bias, such as sex, gender, or racial bias, is a recurring issue in healthcare. For long years, the majority of clinical trials were carried out only in male patients, for reasons such as complex physiology of women due to menstrual cycle [11] or due to concerns regarding the development of the fetus [12]. Furthermore, even when women are included in the clinical trials, sex- or gender-based data are not analyzed separately [12]. As a result, the differences in the manifestation of a given disease in females were not well known. For instance, women are less likely to get correctly diagnosed for having a heart attack because their symptoms can differ from those of male patients [13]. Another example is chronic pain, which affects women more frequently [14, 15]. Yet, it was shown that they receive less painkillers when

they visit the ER and wait longer to receive them compared to men [16]. Even when a disease is more common in women, the research data does not always follow the gender ratios. Although thyroid disorders are more common in women [17], one study investigating effects of high thyroid levels on brain functions recruited only male participants [18]. Additionally, perceived gender identity might have a role in bias in clinical practitioners, and this could affect many areas of clinical process such as delay in diagnosis. One study regarding Parkinson's disease showed that although there was no difference in the time from symptom onset to first physician visit between male and female patients, there was a difference in the duration from first symptom to movement disorder specialist visit (61% greater for women), meaning that gender differences affected the referrals to the correct specialists [19]. So, sex and gender bias exists in various areas of medicine in different forms. A particular point of attention are sex differences in the human brain. Previous studies showed differences in pre-frontal areas, the superior temporal sulcus, the posterior insula, and orbitofrontal cortex, as well as in subcortical temporal structures, such as the amygdala, hippocampus, visual primary cortex, and motor areas [20]. Even after matching for gray matter volume, as the head size differs between males and females, studies showed that sex can still be successfully classified using resting-state brain connectivity measures [21, 22]. Studying sex- and gender-related differences is also important for developing strategies regarding prevention and therapy. For instance, research investigating the effect of physical activity in relation to brain health showed physical activity resulted in greater improvement of cognitive functions in females compared to males [23].

Bias in healthcare is not limited to sex and gender. A meta-analysis covering 2511 studies found that 81% of participants in genome-wide association studies has European ancestry [24]. In a neuroimaging study, differences were found across various brain regions between African-American and white participants [25]. Moreover, although African-American patients are affected by Alzheimer's disease (AD) more often, they are

still underrepresented in studies [26, 27]. Sometimes the data is not disaggregated according to different sex, age, and racial and social groups, and as a result, it is not possible to study the problems of underrepresentation. For example, the frequently used brain atlas, MNI152 template, does not specify the female/male ratio in the population sample used to construct the template [28]. The effects of bias in models used in healthcare are not well understood or well documented. Understanding and documenting the effects of bias is crucial with the recent efforts to collect big data in medicine and recent advances in clinical use of ML models.

Sex and Gender Bias in Machine Learning Models

Algorithms are traditionally defined as a procedure for computers that defines step by step what they should do with a set of input to create an output. By this definition, algorithms are bound to work within the limits of what has been programmed. The need to explicitly program every output quickly becomes a limitation with the growing number of inputs and outputs. To overcome this limitation, ML was born as a field within computer science. ML can be defined as the ability of the algorithm to adaptively learn the outputs from the inputs and to generalize this knowledge on unseen inputs.

Today, ML models play an important role in automating the decision-making process, and society increasingly relies on ML models for supporting decision-making with potentially severe consequences in many areas of human life: from credit scoring and CV screening for job applications to deciding whether defendants awaiting trial should be detained or released [29, 30]. As the algorithms used in automated decision-making are trained using historical data, the learning depends particularly on the so-called labels of training data, which is influenced by the human decision-making process. Human decision-making is and has been affected by various forms of cognitive bias [31], and the data collected regarding those decisions are also prone to bias. As a result,

with the increasing use of big data and ML algorithms for decision-making, algorithmic bias is a growing concern [32]. There are many different underlying reasons of algorithmic bias. One reason is underrepresentation of minorities and lack of diversity in the data and labels collected for training the algorithms. For instance, a recent study has shown that various commercial facial recognition software platforms have significantly reduced performance when applied to women, and the performance difference was even stronger in the case of black women compared to white men [33]. Another example showed that when used for translating job titles from gender-neutral languages to languages with gendered job titles, machine translation uses male defaults for STEM (science, technology, engineering, and mathematics) job titles [34]. Even word embeddings, which have a range of applications from web search results to parsing resumes, show gender bias [35]. Furthermore, the existing biases in the dataset are amplified through the algorithms used. A recent study investigating captioning of the gender in the context of activities seen in an image (for instance, cooking) found that not only the datasets used for training such tasks contained bias, but also the models trained on these data amplified the existing bias [9]. Most importantly, as most of those algorithms are perceived as a black box, the people affected by the algorithmic bias might not be aware of it.

Recently there has been an increasing awareness in the ML community about algorithmic bias. Recent research developed methods about understanding and mitigating algorithmic bias as well as pointing out the importance of algorithmic fairness. The open source Python toolkit AI Fairness 360 (AIF360) is one of the recent tools that provides methods to investigate and mitigate bias in algorithms [10]. Gebru and colleagues proposed datasheets for datasets, where every dataset is accompanied by a datasheet explaining characteristics and recommended use of the dataset [36]. A similar approach, model cards for model reporting, was proposed for documenting dataset characteristics, benchmarking results across different demographic groups and explaining intended use of ML models [37].

Role of Sex and Gender in Machine Learning Models for Medicine

Use of ML models in medicine is gaining popularity especially with increased availability of big datasets and novel algorithms. Thus, understanding the role of sex and gender traits in these models and datasets is becoming more prominent. Medical imaging is one of the cornerstones of diagnostic methods, and with the currently available big datasets, it is perfect for so-called supervised learning methods of traditional ML algorithms as well as convolutional neural networks (CNN), a type of deep learning model [38]. Traditional ML algorithms and statistical methods require domain expertise and manually picking features used as input data. However, deep learning algorithms are data driven, meaning the feature selection can also be automated [39]. Deep learning algorithms can also help finding subtle and complex differences in the data. As a result, deep learning methods are used in various applications in medical imaging, from disease classification for diagnosis to prediction of disease development and treatment outcome [39]. But nevertheless, the need for tuning of hyperparameters and the fact that the deep learning algorithms are perceived as a black box are raising concerns about the fairness of the models across different traits such as sex, gender, and race.

Since the historical data are now being used for training ML algorithms to aid medical decision-making, the lack of diversity and underrepresentation issues are transferred to the models, thus causing algorithmic bias across demographic traits. For instance, a recent study found that a commonly used commercial healthcare risk score prediction algorithm in the United States has racial bias; at a given score, it is showing black patients sicker than white patients [40]. Many neuroimaging studies exclude left-handed participants, thus resulting in a representation bias [41]. When the diversity in the datasets is ignored, the sample distribution would not be similar to the general population, which can influence results [42].

There are still no standardized methods for evaluating models for fairness and bias, and

many models in healthcare are still not reporting results stratified by sex or gender. In a recent review, it was shown that research articles about medical imaging-based diagnosis models rarely report the demographic traits of the dataset used or include the demographic traits as variables in the model [43]. However, one study has indeed demonstrated that sex imbalance in medical imaging (X-ray) datasets used for computer-aided diagnosis (CAD) based on state-of-the-art CNNs produced biased classification results [44]. When they investigated different proportions of sex in the training data, they found that a model trained with balanced and diverse dataset performed best for both male and female patients. Moreover, the balanced dataset resulted not only in an improved performance for the underrepresented group but did also not change the performance for the overrepresented group, so the balanced dataset resulted in an overall better performance [44].

Current Issues

Today, ML is used increasingly in healthcare, and investigation of algorithmic bias is important in order to advance health equity [45]. Gianfrancesco and colleagues investigated the types of bias in ML models trained on electronic health records and suggested paying extra attention to data being used to train the algorithms and making the intended use case of the algorithms clear [46]. Some reasons of bias in healthcare include underrepresentation of female patients [13], lack of diversity [24], and implicit bias by medical professionals [16]. While new tools are developed for the assessment of bias in clinical data and clinical accuracy recently [47, 48], these tools are not specific for machine learning applications, and some of the metrics rely on judgment instead of quantitative measures. More recently, Rajkomar and colleagues categorized key biases in the design, data, and deployment of ML models used in healthcare [45]. For instance, systematical misdiagnosis of a subgroup, differences in the distribution of data across subgroups, features missing not at random in a subgroup as well as clinicians trusting the model even when it is wrong or otherwise ignoring

the outcome of the model, and unavailability of the models to a certain subgroup will all result in unfair outcomes of a ML model [45].

Although deep neural networks are great at modeling the underlying complex and nonlinear characteristics of the data, they are still prone to modeling the confounders that do not have a causal relationship with the outcome but nevertheless can be used to make outcome prediction. The bias in the dataset caused by these confounders will not only be captured but will also be amplified within deep neural networks [42, 49]. Even though this concept is well understood for linear models, it is not so straightforward when it comes to deep neural networks. One example is a skin lesion classification model recognizing the presence of ruler in the skin images as malignant, as the dermatologists would use the ruler when the lesion is more concerning [50].

Dataset bias includes any bias happening in the dataset, possibly due to data collection issues. Lack of diversity, missing data not at random, underrepresentation of certain groups, and transfer of current societal biases via data are some examples of dataset biases [51]. A recent study on prediction of acute kidney injury reported that only 6.38% of the patients were female; thus, the model performance was lower in females [52]. Another type of bias, namely, task bias, can be defined as the intrinsic dependency between a demographic trait and the outcome [53]. For instance, difference in age range between healthy control group and disease group during disease classification task can be given as an example to task bias [53].

Outlook and Potential Solutions

To understand and mitigate bias in ML models in healthcare, the following approaches based on the location of interference in the ML model lifecycle exist (see Table 1):

Collecting Balanced Datasets

Collecting representative and balanced datasets is the most straightforward solution for mitigating bias in ML models. Including both male and

Table 1 Bias detection and mitigation methods grouped based on their respective location in the ML model lifecycle stage

What	When
Collecting balanced datasets	Before model development
Using fairness aware models	During model development
Incorporating fairness metrics	
Measuring bias in model	After model development
Reporting dataset characteristics	
Describing model characteristics	
External auditing of models	
Continuous monitoring of the model performance	

female patients in a balanced way as well as other gender-diverse individuals in any healthcare-related dataset collection process would ultimately mean that the model will learn variability of individuals across sex and gender traits. However, as the data collection is neither easy nor cheap, current use of historical data is still very valuable. Moreover, another issue is that the true distribution of sex- or gender-based disease characteristics on the population level is not always well known. Still, understanding the dataset characteristics based on the demographic traits and reporting those are good practices. Lastly, current initiatives for collecting big data in medicine usually have goals around certain diseases, such as Alzheimer's disease, and those datasets might not be representative of the general population. Benchmark datasets such as ImageNet in computer vision [54] are necessary to compare the performance of machine learning models and are mostly missing in healthcare domain. This is also bringing reliability issues (such as having a significant effect of scanning location in MRI datasets) and bias into ML models in healthcare. Finally, it is important to design data collection with bias awareness in the first place.

Disaggregating Data

As current practices on reporting model performance metrics are not standardized, not reporting on disaggregated data is one of the issues resulting

in not knowing how much bias is included in the machine learning models used for healthcare. In some cases, providing sex-disaggregated data would show opposite effects of a medication across sexes [2]. For instance, levels of hemoglobin A1c (HbA1c) is used for monitoring diabetes, although it was shown to have ethnicity- and gender-specific differences [55, 56]. Here, investigating model performance per subgroups would provide insight about such complex relationships between the bias parameter and model outcome. Standardized reporting on disaggregated data according to demographic traits would help understanding fairness of the ML models developed and increase transparency and trust in the systems designed.

Data and Model Documentation

Good documentation practices help understanding datasets and models and in particular their limitations. The first step would be understanding the limitations of the training dataset and specifics of the model. For the purpose of understanding datasets, Gebru and colleagues developed datasheets for datasets, a documentation model for explaining the data used for training the ML model [36]. A similar documentation method was also developed for the ML models; "model cards for models" was developed as a one-page outline to report standardized results regarding the model performance [37]. For ML models used in medicine, model documentations should help collaboration between clinicians and machine learning researchers [57].

Fairness Aware Models

There is a growing concern regarding fairness in ML models developed and used in everyday life. Various definitions of fairness are introduced based on the final goal. The choice of loss metric of a model is influenced by the fairness metric chosen. Fairness can be on individual level or group level. Individual-level fairness is concerned by having similar treatment for similar individuals, whereas group-level fairness is interested in treatment of different subgroups based on protected parameters such as sex, gender, or race

and obtaining similar model performance across those subgroups. As an example, the performance of a disease prediction model can be evaluated separately on male and female patients using true positive rate for both groups separately. It is important to note that individual fairness metrics might not always be the best fit for using in ML models in medicine. For instance, sex could be affecting the probability of getting an illness; thus, aiming for similar outcome for similar individuals might hinder the model performance. In contrast, the group fairness in this case would aim for equally minimizing misdiagnosis rate across sexes. A study investigating model constraints to match reimbursement for different demographic groups in American healthcare insurance scoring models successfully improved undercompensation for groups without big effects on overall model performance [58].

Especially deep learning networks offer new ways of dealing with bias in the data. One promising approach is adversarial optimization for removing the effects of bias such as sex, gender, or racial bias. Here, while training the deep neural network for the classification task, the model simultaneously disentangles the bias variables by minimizing their classification accuracy on an adversarial network [53, 59].

Although removing bias and correlated features is investigated as a solution for unwanted bias, if the bias is inherently correlated with the disease outcome, such as sex differences in prevalence, removing it could cause unfair outcomes for the minority group [2]. Thus, it is important to understand the relationship between bias and the predicted outcome.

External Auditing Framework for Machine Learning Models

Auditing and regulating models by external bodies is not a standard approach in healthcare ML models as in some other fields such as aviation [51]. Model auditing would inspect the model performance on groups and outcomes post-deployment [51]. Audit studies could document disparities of the models developed [40]. Although the topic of external auditing is still not well

studied in the context of ML models in healthcare, a recent study proposed a framework called FairLens to investigate bias in the decisions of commercial black-box models used in clinical settings [60]. As the ML model use increases in medicine, there will be more need for external auditing and regulation bodies especially when deciding which model benefits which subgroups best. Such comparisons are only reasonable through objective auditing processes.

Continuous Monitoring of the Models

In addition to training ML models, the question of how the trained models perform on new data samples after training and how they generalize to unseen data samples remains another challenge. It is possible that the new data is collected using a new threshold, a different lab instrument, or a different model of MRI scanner. Additionally, if the diversity of the population is not captured in the training data, it is possible that the model will encounter unseen samples belonging to certain racial or gender minorities or different age groups of patients with unseen comorbidities. Continuous monitoring and maintenance of the models as well as taking dataset shift into consideration is vital to keep the model performance from deteriorating. Dataset shift can happen due to many factors such as change in data collection tools and methods, adding data from other sites, policy change, and shift in population [57]. Different approaches to handle dataset shift have been developed recently [61]. However, those approaches usually require predetermination of likely shifts [61]. Alternatively, continuous monitoring of the models and retraining when the performance worsens are recommended [51]. Having the continuous monitoring as part of the ML model development cycle and embedding it into the practices of model development as well as automating the monitoring itself will help to guarantee high quality of ML models in healthcare.

As the use of ML models is still in its infancy in medicine, there are still many obstacles to overcome. ML methods in medicine have a huge potential to provide new personalized diagnostic methods and novel insights into various illnesses.

Standardizing fairness practices will help minimizing gender and sex bias in these methods and help equalize the benefits brought to all gender groups in the society.

References

- WHO. WHO gender policy: integrating gender perspectives in the work of WHO. World Health Organization [Internet]; 2002. <http://apps.who.int/iris/bitstream/10665/67649/1/a78322.pdf>
- Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, Mellino S, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *npj Digit Med* [Internet]. 2020;3(1):81. <https://doi.org/10.1038/s41746-020-0288-5>.
- Reisner SL, Conron KJ, Tardiff LA, Jarvi S, Gordon AR, Austin SB. Monitoring the health of transgender and other gender minority populations: validity of natal sex and gender identity survey items in a U.S. National cohort of young adults. *BMC Public Health*. 2014;14(1):1–10.
- Legato MJ. Rethinking gender-specific medicine. *Women's Health*. 2006;2(5):699–703.
- de Paredes ES. Gender bias in medicine. *Appl Radiol*. 2004;33(9):6.
- Ngo ST, Steyn FJ, McCombe PA. Gender differences in autoimmune disease. *Front Neuroendocrinol* [Internet]. 2014;35(3):347–69. <https://doi.org/10.1016/j.yfrne.2014.04.004>.
- Maas AHEM, Appelman YEA. Gender differences in coronary heart disease. *Neth Heart J*. 2010;18(12):598–603.
- Östman J, Lönnberg G, Arnqvist HJ, Blohmé G, Bolinder J, Schnell AE, et al. Gender differences and temporal variation in the incidence of type 1 diabetes: results of 8012 cases in the Nationwide Diabetes Incidence Study in Sweden 1983–2002. *J Intern Med*. 2008;263(4):386–94.
- Zhao J, Wang T, Yatskar M, Ordóñez V, Chang KW. Men also like shopping: reducing gender bias amplification using corpus-level constraints. *EMNLP 2017 – Conf Empir Methods Nat Lang Process Proc*. 2017;2979–89.
- Bellamy RKE, Mojsilovic A, Nagar S, Ramamurthy KN, Richards J, Saha D, et al. AI fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. *IBM J Res Dev*. 2019;63(4–5):4:1.
- Verdonk P, Benschop YWM, De Haes HCJM, Lagro-Janssen TLM. From gender bias to gender awareness in medical education. *Adv Heal Sci Educ*. 2009;14(1):135–52.
- Howard LM, Ehrlich AM, Gamlen F, Oram S. Gender-neutral mental health research is sex and gender biased. *Lancet Psychiatry* [Internet]. 2017;4(1):9–11. [https://doi.org/10.1016/S2215-0366\(16\)30209-7](https://doi.org/10.1016/S2215-0366(16)30209-7).
- Mosca L, Banks CL, Benjamin EJ, Berra K, Bushnell C, Dolor RJ, et al. Evidence-based guidelines for cardiovascular disease prevention in women: 2007 update. *Circulation*. 2007;115(11):1481–501.
- Breivik H, Collett B, Ventafridda V, Cohen R, Gallacher D. Survey of chronic pain in Europe: prevalence, impact on daily life, and treatment. *Eur J Pain* [Internet]. 2006;10(4):287. <https://doi.org/10.1016/j.ejpain.2005.06.009>.
- Tsang A, Von Korff M, Lee S, Alonso J, Karam E, Angermeyer MC, et al. Common chronic pain conditions in developed and developing countries: gender and age differences and comorbidity with depression-anxiety disorders. *J Pain*. 2008;9(10):883–91.
- Chen EH, Shofer FS, Dean AJ, Hollander JE, Baxt WG, Robey JL, et al. Gender disparity in analgesic treatment of emergency department patients with acute abdominal pain. *Acad Emerg Med*. 2008;15(5):414–8.
- Bauer M, Glenn T, Pilhatsch M, Pfennig A, Whybrow PC. Gender differences in thyroid system function: relevance to bipolar disorder and its treatment. *Bipolar Disord*. 2014;16(1):58–71.
- Göbel A, Heldmann M, Göttlich M, Dirk AL, Brabant G, Münte TF. Effect of mild thyrotoxicosis on performance and brain activations in a working memory task. *PLoS One*. 2016;11(8):1–15.
- Saunders-Pullman R, Wang C, Stanley K, Bressman SB. Diagnosis and referral delay in women with Parkinson's disease. *Gend Med* [Internet]. 2011;8(3):209–17. <https://linkinghub.elsevier.com/retrieve/pii/S155085791100074X>
- Lotze M, Domin M, Gerlach FH, Gaser C, Lueders E, Schmidt CO, et al. Novel findings from 2,838 adult brains on sex differences in gray matter brain volume. *Sci Rep* [Internet]. 2019;9(1):1–7. <https://doi.org/10.1038/s41598-018-38239-2>.
- Dhamala E, Jamison KW, Sabuncu MR, Kuceyeski A. Sex classification using long-range temporal dependence of resting-state functional MRI time series. *bioRxiv* [Internet]. 2019. <https://doi.org/10.1101/809954>.
- Weis S, Patil KR, Hoffstaedter F, Nostro A, Yeo BTT, Eickhoff SB. Sex classification by resting state brain connectivity. *Cereb Cortex*. 2020;30(2):824–35. <https://doi.org/10.1093/cercor/bhz129>
- Barha CK, Hsu CL, ten Brinke L, Liu-Ambrose T. Biological sex: a potential moderator of physical activity efficacy on brain health. *Front Aging Neurosci*. 2019;11:1–10.
- Popejoy AB, Fullerton SM. Genomics is failing on diversity. *Nature*. 2016;538(7624):161–4.
- Isamah N, Faison W, Payne ME, MacFall J, Steffens DC, Beyer JL, et al. Variability in frontotemporal brain structure: the importance of recruitment of African Americans in neuroscience research. *PLoS One*. 2010;5(10):1–6.
- Gilmore-Bykovsky AL, Jin Y, Gleason C, Flowers-Benton S, Block LM, Dilworth-Anderson P, et al. Recruitment and retention of underrepresented populations in Alzheimer's disease research: a systematic review. *Alzheimer's Dement Transl Res Clin Interv* [Internet]. 2019;5:751–70. <https://doi.org/10.1016/j.trci.2019.09.018>.

27. McDonough IM. Beta-amyloid and cortical thickness reveal racial disparities in preclinical Alzheimer's disease. *NeuroImage Clin* [Internet]. 2017;16:659–67. <https://doi.org/10.1016/j.nicl.2017.09.014>.
28. Grabner G, Janke AL, Budge MM, Smith D, Pruessner J, Collins DL. Symmetric atlasing and model based segmentation: an application to the hippocampus in older adults. *Lect Notes Comput Sci* (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2006;4191 LNCS:58–66.
29. Zliobaite I. Measuring discrimination in algorithmic decision making. *Data Min Knowl Discov*. 2017;31: 1060–89.
30. Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A. Algorithmic decision making and the cost of fairness. *Proc ACM SIGKDD Int Conf Knowl Discov Data Min*. 2017;Part F1296:797–806.
31. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. *Science* (80-) [Internet]. 1974;185(4157):1124–31. c:%5CICT%5CEILS%5CHyperbole Systeme%5C1973 Rep Effect of Pressure on Ignition of Hypergolic Liquid Propellants.pdf TS – RIS.
32. Hajian S, Bonchi F, Castillo C. Algorithmic bias: from discrimination discovery to fairness-aware data mining. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining – KDD '16 [Internet]. New York: ACM Press; 2016. p. 2125–6. <http://dl.acm.org/citation.cfm?doid=2939672.2945386>.
33. Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. In: Friedler SA, Wilson C, editors. Proceedings of the 1st conference on fairness, accountability and transparency [Internet]. New York: PMLR; 2018. p. 77–91. (Proceedings of Machine Learning Research; vol. 81). <http://proceedings.mlr.press/v81/buolamwini18a.html>.
34. Prates MOR, Avelar PH, Lamb LC. Assessing gender bias in machine translation: a case study with Google Translate. *Neural Comput Appl*. 2020;32:6363–81. <https://doi.org/10.1007/s00521-019-04144-6>
35. Bolukbasi T, Chang KW, Zou J, Saligrama V, Kalai A. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Adv Neural Inf Process Syst*. 2016;29:4356–64.
36. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé H, et al. Datasheets for datasets. 2018. <http://arxiv.org/abs/1803.09010>
37. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model cards for model reporting. *FAT* 2019 – Proc 2019 Conf Fairness, Accountability, Transpar*. 2019;220–9.
38. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42: 60–88.
39. Vieira S, Pinaya WHL, Mechelli A. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications. *Neurosci Biobehav Rev* [Internet]. 2017;74:58–75. <https://doi.org/10.1016/j.neubiorev.2017.01.002>.
40. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* (80-). 2019;366(6464):447–53.
41. Bailey LM, McMillan LE, Newman AJ. A sinister subject: quantifying handedness-based recruitment biases in current neuroimaging research. *Eur J Neurosci*. 2019;51:1642–56.
42. Lewinn KZ, Sheridan MA, Keyes KM, Hamilton A, McLaughlin KA. Sample composition alters associations between age and brain structure. *Nat Commun* [Internet]. 2017;8(1). <https://doi.org/10.1038/s41467-017-00908-7>.
43. Abbasi-Sureshjani S, Raumanns R, Michels BEJ, Schouten G, Cheplygina V. Risk of training diagnostic algorithms on data with demographic bias. *arXiv Prepr arXiv* [Internet]. 2020;20:1–9. <http://arxiv.org/abs/2005.10050>
44. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc Natl Acad Sci* [Internet]. 2020;201919012. <http://www.pnas.org/lookup/doi/10.1073/pnas.1919012117>
45. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med*. 2018;169(12): 866–72.
46. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med* [Internet]. 2018;178(11):1544. <http://archinte.jamanetwork.com/article.aspx?doi=10.1001/jamainternmed.2018.3763>
47. Whiting PF. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* [Internet]. 2011;155(8):529. <http://annals.org/article.aspx?doi=10.7326/0003-4819-155-8-201110180-00009>
48. Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Ann Intern Med*. 2019;170(1):W1–33.
49. Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller KR. Unmasking Clever Hans predictors and assessing what machines really learn. *Nat Commun*. 2019;10(1):1096.
50. Narla A, Kuprel B, Sarin K, Novoa R, Ko J. Automated classification of skin lesions: from pixels to practice. *J Invest Dermatol* [Internet]. 2018;138(10):2108–10. <https://doi.org/10.1016/j.jid.2018.06.175>.
51. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in health. 2020;1–24. <http://arxiv.org/abs/2009.10576>
52. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute

- kidney injury. *Nature* [Internet]. 2019;572(7767):116–9. <https://doi.org/10.1038/s41586-019-1390-1>.
53. Adeli E, Zhao Q, Pfefferbaum A, Sullivan EV, Fei-Fei L, Niebles JC, et al. Representation learning with statistical independence to mitigate bias. 2019. <http://arxiv.org/abs/1910.03676>
54. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis.* 2015;115(3):211–52.
55. Bae JC, Suh S, Jin SM, Kim SW, Hur KY, Kim JH, et al. Hemoglobin A1c values are affected by hemoglobin level and gender in non-anemic Koreans. *J Diabetes Investig.* 2014;5(1):60–5.
56. Cavagnoli G, Pimentel AL, Freitas PAC, Gross JL, Camargo JL. Effect of ethnicity on HbA1c levels in individuals without diabetes: systematic review and meta-analysis. *PLoS One.* 2017;12(2):1–14.
57. Saleh S, Boag W, Erdman L, Naumann T. Clinical collabsheets: 53 questions to guide a clinical collaboration. *Proc Mach Learn Res.* 2020; (MI):1–29.
58. Zink A, Rose S. Fair regression for health care spending. *Biometrics.* 2020;76(3):973–82.
59. Zhang BH, Lemoine B, Mitchell M. Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society [Internet]. New York: ACM; 2018. p.335–40. <https://dl.acm.org/doi/10.1145/3278721.3278779>
60. Panigutti C, Perotti A, Panisson A, Bajardi P, Pedreschi D. FairLens: auditing Black-box clinical decision support systems. 2020. <http://arxiv.org/abs/2011.04049>
61. Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics.* 2020;21(2):345–52.



Meta Learning and the AI Learning Process

29

Samyakh Tukra, Niklas Lidströmer, and Hutan Ashrafiyan

Contents

Introduction	408
General Considerations	409
Transfer Learning	410
Few-Shot Learning	411
Continual Learning	412
Multi-task Learning	413
Neural Architecture Search	416
Conclusion	418
References	418

Abstract

The huge torrent in data collection and the advent of deep learning have been transformative in artificial intelligence research, where deep learning models have achieved enormous success in divergent complex tasks ranging from computer vision to robotic control. However, the success of these models necessitates large quantities of data and exhaustive computational resources. Meta learning on the contrary aims to improve the AI learning process by imitating the human learning process, thereby enabling the AI to learn new concepts and generalize even from few samples of data. In this chapter, we highlight the most prominent approaches in meta learning and its applicability in medicine.

S. Tukra (✉)
Department of Surgery and Cancer, Imperial College
London, London, UK
e-mail: samyakh.tukra17@imperial.ac.uk;
samtukra@thirdeye.health

N. Lidströmer
Department of Women's and Children's Health, Karolinska
Institutet, Stockholm, Sweden
e-mail: niklas.lidstromer@ki.se; niklas@lidstromer.com

H. Ashrafiyan
Department of Surgery and Cancer, Imperial College
London NHS Trust, London, UK
Institute of Global Health Innovation, Imperial College
London, London, UK

Hamlyn Centre for Robotics and Artificial Intelligence,
Department of Surgery and Cancer, Imperial College
London, London, UK
e-mail: h.ashrafiyan@imperial.ac.uk

Keywords

Meta learning · Learning to learn · Few-shot learning · Transfer learning · Supervised learning · Reinforcement learning · Continual learning · Neural architecture search · Neural networks · Deep learning · Machine learning · Artificial intelligence

Introduction

Artificial intelligence (AI) is transforming almost every industry due to its versatile nature of computation. It is endowed with the ability to extend and adapt to any data it is presented with and thereby optimize itself to achieve any given task. This feature in particular can be achieved through deep learning. If you imagine AI as a set of Russian nesting dolls, AI is the largest doll encasing the others, followed by machine learning, with deep learning nestled inside. Each of these aforementioned subdivisions is a component of the prior term. Deep learning is the most successful class of machine learning algorithms, and it is comprised of artificial neural networks. These neural networks loosely simulate the human brain, whereby adjusting neural connections enables the algorithm to learn and perform complex tasks with increasing accuracy independently (without any human intervention). Similar to the networks in the brain, deep learning models are multilayered, which allows them to learn multiple levels of abstractions of the underlying data.

Deep learning-based approaches have achieved great success in a variety of fields ranging from computer vision, natural language processing, to autonomous control in robotics and much more. Specifically, in medicine, it has dominated in the realm of disease detection and suggesting treatments. However, there are limitations to where AI can be applied. Thus far, successful algorithms have been largely dependent on vast quantities of data and where exhaustive computational resources are available, for the training to be conducted over a long duration. This poses a huge problem for further development of AI models, since it may not always be possible to collect large quantities of data for a

specific task nor have the resources to train large models for an extensive amount of time, such as GPT-3 [1]. These settings break the current AI paradigm, as it is not feasible for an AI model to learn every single new task from the outset (which would also require large amounts of data for support). This is exactly the problem **meta learning** (also known as “learning to learn”) aims to solve. It is an exciting research direction that aims to strengthen the approach of training AI models, similar to how learning in the human brain occurs. To understand this further, one can draw parallels between machine learning and human learning.

Humans do not just learn new concepts; they also learn associated biases so that they can learn to generalize. This enables humans to use the lessons gained from past experiences or different tasks as a foundation to learn new ones. As a result, humans are able to generalize correctly from very few data samples. Sometimes it only takes a single sample/experience to learn a new skill [2]. Human learning is selective; it is able to re-utilize approaches that worked well in previous experiences and focus on enhancing them in new challenges. This makes acquiring new skills easier and less dependent on multiple attempts to learn. The research question that meta learning aims to answer is: *Is it possible to design machine learning models capable of adapting and generalizing to new tasks that have not been encountered before, with relatively low training data, similar to how humans adapt?* Understanding the behavior of machine learning techniques in such challenging and extremely difficult tasks can enable the design of processes to help the model make the correct choices.

Therefore, meta learning is also known as *learning to learn*, as it defines a set of methods that can be used to rethink the AI training process. A deeper understanding of computer programs that focuses on the improvement of their ability to learn new concepts faster and more effectively can have enormous practical impact in the field of medicine. Examples include working with electronic health records (EHRs) [3]. These EHRs contain the necessary data to conduct health data analytics such as prediction of in-hospital mortality, re-admission, disease classification, and many more. However, the data contained in an EHR

poses its own challenges like irregularity (e.g., irregular time series signals), sparsity (missing data), imbalance in labels, and many more. Hence the authors used meta learning for risk prediction and disease classification, where a “meta learner” model was trained entitled “Meta-Pred” that learns the domain knowledge for deep feature learning, which is later fine-tuned via transfer learning for specific tasks like disease classification, etc.

Meta learning has also extended to:

- Medical image segmentation [4], where the authors utilized meta learning to evaluate and select models for segmentation efficiently by predicting their performance for new tasks
- Multi-modality medical image segmentation, achieved using a single convolutional neural network via multi-task learning in [5]
- Skin disease identification, addressed by meta learning (few-shot learning in particular) in [6]
- Dermatological disease identification, a huge problem where deep networks must be capable of adapting to novel diseases while being robust to changes in images such as lighting, brightness, color tones, etc.

Meta learning generally enables deep learning models to adapt to the ever-changing new environments, thus making them ideal for exploration in medical applications.

This chapter highlights a survey of the frontiers of the most prominent and exciting approaches in meta learning, with the objective being to investigate the feasibility of such methods applied in medical AI research to enhance the development of AI solutions. In particular, the general foundation of meta learning will be reviewed and compared with regular machine learning paradigms. Furthermore, specific methods for improving data representation in meta learning will be discussed by exploring the following sections: “[Transfer Learning](#),” “[Few-Shot Learning](#),” and “[Continual Learning](#).” Next, methods specific to making deep learning models efficient at meta learning will be discussed via neural architecture search. Finally, an outlook into where the general direction of future research is heading to will be discussed.

General Considerations

Prior to delving into the specifics of meta learning research, it is vital to have a solid foundation for investigating this notion further. In this section, the general areas of machine learning will be discussed, followed by its contrast with meta learning and how the two are related. Consider in particular two different paradigms of machine learning: supervised learning and reinforcement learning. In **supervised learning** the machine learning model learns from a set of ground truth values captured in the data. For example, a dataset comprising of a set of images with and without skin lesions is provided, including a set of labels associated with these images defining their class. The task of the model is to recognize and correctly classify the respective image. The model optimizes for predicting the correct labels, but any task beyond what is represented in this dataset will result in poor performance. The model would require to be trained once again from the beginning. With meta learning the aim would be to use the knowledge gained from this training and apply it to other tasks and datasets, instead of having to train again from the outset. To achieve this, a meta-training dataset is defined, which is a set of multiple datasets. These multiple datasets are for different tasks, though they must have structural similarity (i.e., still visual recognition, but not identical to the original). The aim is for the model to learn meta-knowledge, so that it can learn a variety of different tasks well; hence it is “learning to learn.” However, retaining this meta-training dataset forever in memory is inefficient, as it requires storing large amounts of distinct datasets. Instead, in addition to training the base AI model (referred to as the “base-learner”), a “meta learner” model is learned. This “meta learner” model learns all the information required for solving new tasks from the “meta-training” data. The objective of meta learning is thereby solving or estimating the best possible “meta learner” model (usually another deep learning model). The “base-learner” then uses the “meta learner” to hone its predictions to get high performance, known as the adaptation process, illustrated in Fig. 1. Hence, meta learning is broken down into two phases: (1) meta-training phase and (2) adaptation phase.

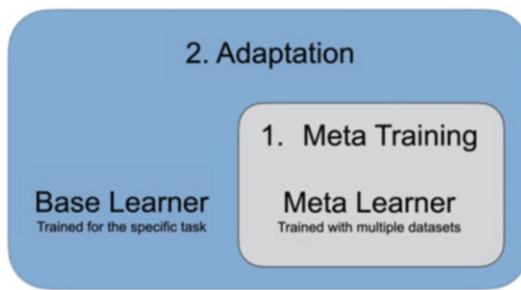


Fig. 1 Displaying the supervised meta learning process. Comprising of two training steps: 1. meta-training and 2. adaptation

In **reinforcement learning** the “base-learner” model learns from experience. Unlike supervised learning where ground truth labels are available, reinforcement learning enables the model to learn from trial-and-error. Consider an example of playing a game. This problem comprises of an environment within the game and the character the AI model has to control. This character is referred to as the agent. The aim of the agent is to traverse through the game such that it always wins. Hence the model must select the best actions conducted by the agent to always win the game. The definition of winning is defined by how much reward is earned by the agent (dictated by a mathematical reward function). The higher the reward, the better the selected action and therefore the higher the likelihood of winning. *The reinforcement learning goal is where the agent learns how to act in order to maximize its expected reward.* However, training reinforcement learning algorithms, despite being powerful, have major challenges, such as long training times [7, 8], lack of multi-tasking, and algorithmic training instability.

Given that meta learning aims to solve generalizability to different tasks, this paradigm can be re-factored into a meta learning process, i.e., **meta-reinforcement learning** to enable generating more efficient reinforcement learning algorithms. The ability to learn from multiple data distributions to solve unseen tasks grows extremely broad, enabling AI to be en route toward general-purpose methods. Using the analogy of meta-supervised learning defined above, meta-reinforcement learning can be introduced also as a two-step process. Reinforcement learning does not comprise of data as in supervised;

instead it comprises of a Markov decision process (MDP), a set of state, actions, rewards, and next state pairs. Where state is an instance of an environment at a given time, action is the predicted action the agent performs, and reward is the respective reward attained for that action, and finally next state is an instance of an environment at the next time step. Hence, instead of the “meta learner” training on meta-training data (multiple datasets), it trains on multiple MDPs, i.e., meta-training MDPs, where each MDP is associated with some different task.

Transfer Learning

Transfer learning is the simplest form of meta learning, often conducted in deep learning. The aim of transfer learning is to transfer the knowledge of previous tasks with larger data, learned by some “base-learner,” to new unseen tasks via further training or commonly known as “fine-tuning.” The focus here is only on the adaptation phase, as the assumption is that the meta-training phase already occurred by training on large datasets like in [9]. When the dataset lacks sample diversity or quantity, the deep learning model is unable to extract strong features. Thus, re-training a “base-learner” that has imbibed the meta-knowledge it gained from the original task can enable learning a new one, faster. This is achieved by selecting a state-of-the-art model for a given task together with their corresponding learned parameters, commonly known as “weights.” For example, in a visual recognition task, such models are in [10, 11]. This pre-trained model would be initiated with its weights (i.e., the meta-knowledge) and retrained on the new dataset/task. Since the original parameters are reused to initialize the training process, the knowledge of the previous task is carried forward to learn the new one. This results in both shorter training times due to higher likelihood of convergence and strong features being learned, despite the presence of a small number of data samples. This outcome would not be possible, if the model would have started the training from square one.

This has profound impact in the medical machine learning community due to data scarcity for many problems. Consider the following

example, where skin disease detection was performed using the aforementioned state-of-the-art vision models in [12]. In another study, a similar method was utilized for detection of diabetic retinopathy [13]. Similar cases are applied in different domains of machine learning, like in medical natural language understanding where state-of-the-art natural language processing (NLP) models like [14] were fine-tuned on medical text data to attain cross-domain high performance [15]. However, the biggest limitation of transfer learning is that the benefit of a pre-trained model decreases substantially if the desired target task diverges from the original, as shown in [16]. The key here is to initiate the training process with some prior meta-knowledge. Hence, it is imperative that we define a method to attain a generalizable and a valuable initialization of model weights that serves as a good foundation to initiate further fine-tuning for other datasets/tasks.

Few-Shot Learning

Humans can usually learn to generalize from very small amounts of data; sometimes it just takes a single sample. Consider face recognition: for humans only one image of the face is required to recognize it again, despite variations in environment which may change the way a face appears such as different angles, lighting, brightness, size, orientation, race, etc., yet humans do not need to see the original image sample again. Few-shot learning aims to achieve this similar characteristic, unlike transfer learning which still requires additional data for training. Few-shot learning is also referred to as N-shot learning, where N is the number of samples, i.e., 1-shot learning denotes learning from a single sample, likewise 5-shot learning denotes learning from five samples. There are three approaches to this meta learning paradigm: metric-based, model-based, and optimization-based. Obviously using traditional methods of training on such a handful of (N) samples will result in over-fitting and the model performing poorly. The three approaches aim at alleviating this problem.

Metric-based methods involve learning a distance function over the data samples, similar

to nearest neighbor algorithms like K-NN classifiers [17] or K-means clustering [18]. To comprehend how this works, consider an image as an input, which the model must classify correctly to some label. This input image can be projected to a lower dimensional representation by parsing it through the model. This representation is known as an “embedding,” which acts like compression. A fascinating characteristic of deep learning models is that these output compact representations (embeddings) are mostly unaffected by intra-class variations all while retaining information about class divisions. Images of similar classes will have embeddings clustered together, and those of different classes will be far apart. This is further visualized in Fig. 2. Learning this distance function between an embedding for a new unseen image, with respect to other embeddings in the model’s meta-knowledge, enables classification without re-training/over-fitting.

The simplest form of **metric-based** approach in deep learning is the use of a Siamese twin neural network, comprising of twin input heads. The deep learning model learns the relationship between the two inputs and thereby the distance metric [20]. If the two inputs are substantially different, i.e., from a different class, then this distance is large, and if they are similar, i.e., intra-class, it is small enabling classification with their respective label. Furthermore, Vinyals et al. [21] took this idea further by learning a classifier directly that infers the conditional probability

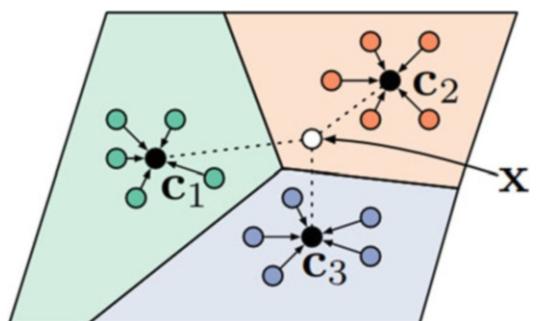


Fig. 2 Displaying embedding space visualization where the embedding of x is being classified among other learned embeddings, where each cluster centroid embedding is shown as c and x being the new sample embedding. (Figure source: [19])

distribution of the different input types itself. Similarly, there have been more advancements in few-shot learning building on top of the aforementioned work like [22] which is similar to the Siamese network with some differences in the training procedure and the architecture itself. [19] directly learns an embedding function to encode each input into a user-defined feature vector referred to as “prototype feature vector.” A distance is then calculated between the prototype clusters and the unlabeled image embeddings to perform few-shot classification.

The **model-based** approach still relies on the notion of embeddings; however it focuses more on the model design process, specifically for fast learning. The objective is to design a model capable of rapid parameter update in just a small number of training steps. This swift parameter update is achieved via architecture design capable of harnessing prior knowledge to reduce training time and size of hypothesis space or via an external meta-learner model. Examples of this type of approach include a family of architectures which use internal memory modules to read from and write to and a learning process which can be later exploited for learning in new cases [23]. This includes [24] which introduces a model and training process specifically designed for rapid generalization, by relying on “fast weights.” This is an interesting approach, but unfortunately there is limited agreement on a concrete definition. However, it can be understood by comparison to traditional techniques. Usually, deep learning models are trained with gradient descent-based algorithms which are slow and take time to converge to the optimal minima. A faster approach would be to utilize another neural network to predict the parameters of the target/candidate neural network. These predicted parameters are then further optimized and known as “fast weights.”

Optimization-based approaches aim to innovate on the methods utilized for the optimization process in deep learning. Currently, deep learning relies on gradient-based optimization, i.e., gradient descent, which struggles for convergence when there is lack of data and when the learning rate is small. The objective thus is to design optimization algorithms such that the models can learn, despite availability of only a few samples

of data. Methods encompassing this type of approach include [25] where the authors propose a long short-term memory (LSTM)-based meta learner network that learns the optimization algorithm itself to train the “learner” neural network for classification. [26] proposed a model-agnostic meta learning algorithm (MAML) that is a general optimization algorithm, which is compatible with any model and any problem, i.e., supervised, unsupervised, reinforcement learning, etc., as it learns through traditional gradient-descent algorithms. Similar to MAML, Reptile was introduced by [27]. It builds on MAML and works by sampling a new task from a dataset repeatedly and moving the initialization of parameters toward the trained weights on that task.

Some examples of few-shot learning applied in the medical domain include segmentation, in particular 3D segmentation of volumetric medical images (CT scans) by [55]. Here the authors proposed their own few-shot learning framework in the context of pixel-wise classification, utilizing only a few annotated slices. [56] experimented with few-shot learning in the context of dermatological disease classification. The authors tried to address the problem of intra-class variability in disease types being too large and the lack of balanced data to support. In particular they extend [19] for clustering. Furthermore, [57] utilized one-shot learning for drug discovery, reducing the necessitating on large quantities of data to make meaningful predictions in such applications. They utilized a long short-term memory-based hybrid architecture combined with graph neural networks to improve the distance metric learning between molecular structures.

Continual Learning

There is still a lot of ambiguity in how humans learn. However, it is certain that humans can continuously learn and accumulate new knowledge and skills to be reused in new problem case scenarios. This ability of retaining knowledge of past experiences for fast adaptation is what continuous learning aims to mimic when training machine learning models. Hence, it is also referred to as “cumulative learning” and “life-long learning,” as

the objective encapsulates around ensuring that the model does not forget its previous knowledge. Hence, the objectives are the following:

1. Accelerate future learning by exploiting previous/existing knowledge regarding the task
2. Avoiding overwriting or interfering with previous knowledge while acquiring new knowledge – this is commonly known as *catastrophic forgetting*

The ability to store knowledge in memory is innate to humans. For example, when a human tries to learn a new language, they do not forget their original tongue to do so; instead they retain the original knowledge while learning new information. However, this limitation is continuously seen in many deep learning systems [28], even in state-of-the-art models, which raises serious concerns regarding scalability of systems in the real world, especially in medicine which is a dynamic environment with multiple changes occurring for which data collection cannot necessarily be planned. Hence, the aim of continual learning is to simply never stop the training process of the model, via an endless stream of data being processed by the model. The model thus has to learn from multiple tasks, based on the knowledge of previous tasks, enabling it to generalize even if it has never observed some future task. This is typically used in self-driving cars [29] and in recommendation systems [30]. The idea is to continually train our prediction models, in small increments as new data samples become available. This aids in **efficiency** of AI models, since training from square one is no longer required hence conserving computational power while ensuring models are up to date. Adabtability is improved when a system trains faster and can guarantee adaptation to new situations using the training methodologies described above like few-shot learning or transfer learning. Lastly, for **scalability**, the memory overhead remains low throughout the model life cycle, since data is not stored but directly learnt from, while processing data in real time.

However, to make it actually work is non-trivial, requiring engineers to balance the *stability-plasticity dilemma*. The stability-plasticity dilemma stems from biological neural networks,

where the so-called synaptic consolidation gives rise to continual learning via reduction of the plasticity of synapses that are vital to previously learned tasks. This enables neuro-synaptic adaptation to changes in the environment and thus allowing it to change the neurons physical structure as a result of this learning. However, for this effect to take place, a balance must be maintained such that episodic memories are retained (memorization) and generalizability is not affected via maintenance of some amount of plasticity, hence the stability-plasticity dilemma. [28] introduced a new method of updating weights via *elastic weight consolidation*. This method of addressing the catastrophic forgetting issue falls into the category of modification of the online update to retain knowledge. There are additional methods of tackling this problem that involve designing the model itself in an innovative fashion. An example is [31] which introduced “progressive neural networks” which are immune to catastrophic forgetting, by progressively growing the neural network with each task. Additionally, networks can be equipped with episodic memory such that task-based features are retained and memorized to be reused in new tasks [32]. Finally, another interesting approach comprises of generating samples of past tasks for more updates via the popular generative adversarial networks framework [33] and via using semi-distributed representations. Due to the key three features continual learning offers, efficiency, adaptiveness, and scalability, this paradigm can be applied to any model that would be potentially deployed in the clinical world. Medical AI models that are production level and to be deployed in real settings can have a huge advantage by exploiting continual learning. Since the models will actively learn from incoming (infinite) data, it will self-update and adapt, enabling a dynamic prediction/decision-making model.

Multi-task Learning

Up till this point, in most of the aforementioned types of learning the tasks defined still belonged to a similar category (e.g., classification). Multi-task learning focuses on endowing the AI model with the capability of conducting multiple tasks that

can be divergent in category, i.e., in the example of computer vision, predicting depth and optical flow (which are two separate tasks). The aim is to design deep learning models or optimization strategies that enable the model to learn shared representations from the multi-task signals. These shared representations provide substantial benefits, such as data efficiency and faster future/transfer learning speeds for future downstream tasks and thus also aiding in alleviating the common aforementioned weaknesses in deep learning.

However, there are complications associated with multi-task learning: since the model has to optimize for multiple tasks, there is a possibility the different tasks may have conflicting feature requirements. Therefore, the model may perform really well on one of the tasks but in turn hurt the performance on the other tasks which require a different distribution of features, a problem commonly known as *negative transfer*. It is vital that such models/optimization techniques balance the flow of information between tasks: yielding positive transfer in the model weights while learning to discourage sharing when the potential of negative transfer rises. The goal thus becomes designing such techniques and is the essence of multi-task learning research.

As stated above, there are two approaches to multi-tasking. First is the architecture design that enables the model to learn both features that should be shared across tasks and features that should be task-specific, such that a generalizable representation is learned. Second is designing the optimization strategy (i.e., the loss function) that is capable of balancing the learning signal across different tasks, providing equal importance to all.

Architecture design must achieve a shared representation learning across tasks. Typically there are two methods of achieving this sharing: (1) hard- and (2) soft-parameter sharing. In hard-parameter sharing, the model is divided into two parts, initiating with a fixed part, i.e., like a backbone that consists of a shared encoder, which learns inter-task features directly from the input. This encoder then branches out into the second part, which are task-specific heads. These heads learn the inter-task features that are specific to the task output itself. Fig. 3 exhibits this visually.

The task-specific heads use the features learned from the backbone, to then learn inter-class-specific features to hone the output. Having a shared backbone reduces the risk of overfitting for one particular task since feature distribution is saturated across them. However, it also raises design concerns, like: where would be the best point to initiate task-specific heads? Or how large should the backbone be? etc. Such concerns are usually alleviated heuristically based on design and experimentation. The objective therefore is to find the best potential configuration of such models. Some of the earliest examples of this method include [34], which utilized a shared backbone followed by task-specific heads for facial landmark detection, joint head pose estimation, and attribute inference.

There were several works following the aforementioned one that proposed variations of this shared backbone design, i.e., [35, 36]. [37] introduced Multi-task Network Cascades, for instance, level segmentation. The architecture is very similar to [34] with the exception that the output of each task-specific head is appended to the input of the next sequential task-specific head, thereby creating the “cascade” of information flow. [38] aimed to jointly tackle a large number of vision tasks via multi-head design. Their model not only featured multi-head design but extends it across different network layers and scales too. In all cases, the network connections between shared backbone and the task-specific heads are heuristically determined via trial-and-error which is marginally suboptimal. Hence, some of the latest works try to overcome this issue by proposing adaptive techniques for automatically designing such networks and perform decision-making for determining the optimal position for sharing information. Such methods include neural architecture search (described further in section “Conclusion”) to design such architectures [39–41].

In contrast, in soft-parameter sharing, each task has its own assigned set of parameters (i.e., individual network architectures), with some method of sharing these features across other task-specific networks. This is illustrated as the dotted lines in (b) in Fig. 3. This method of sharing features across individual task-specific networks is

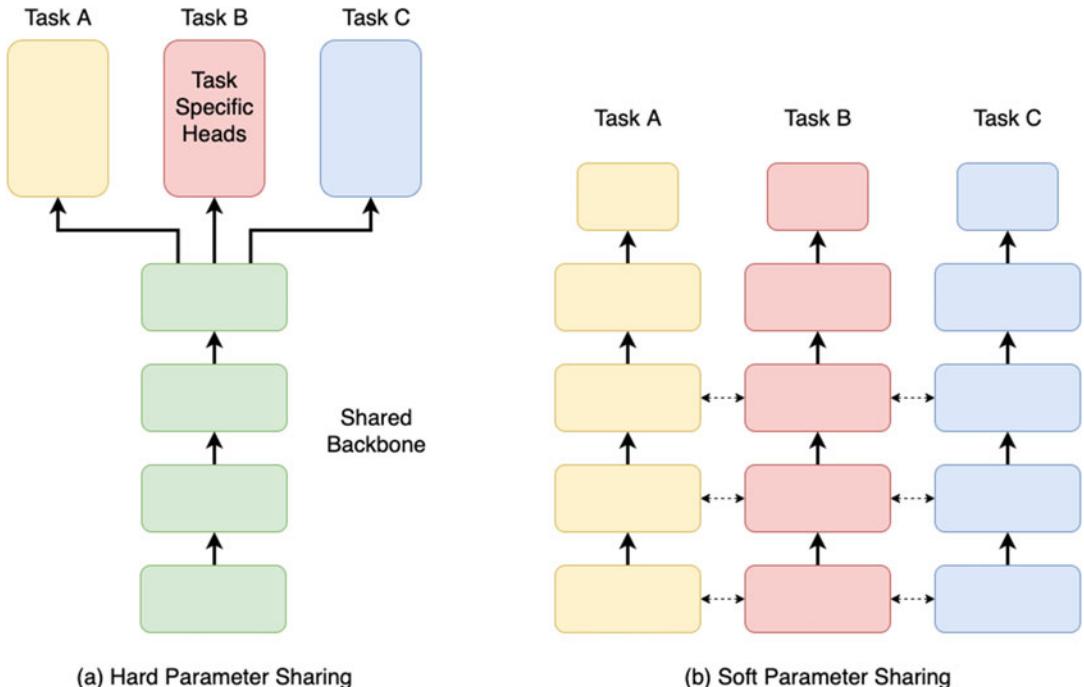


Fig. 3 Exhibiting architecture design for (a) hard- and (b) soft-parameter sharing. (a) A backbone is created as a shared trunk that learns the inter-task features. Followed

by individual task-specific heads, (b) each task has an individual architecture with cross-sharing of features between them

commonly known as “cross-talk,” allowing for information to flow between all layers in parallel. The objective is to find the most effective way of conducting this cross-talk across the task-specific networks. One of the earliest works to achieve this is [42], which introduced “cross-stitch” networks. Here the input of every layer is a linear combination of previous layer outputs. The linear combination weights were optimized for a specific task, such that each layer could decide autonomously which inter-task information to leverage. Later, Sluice Networks were introduced by [54], who adopted the idea of cross-stitch networks and extended it by allowing it to directly learn the selective sharing capability of the layers and shared sub-spaces. [43] further added dimensionality reduction via 1×1 convolution, to the feature fusion components of the network.

From the optimization point of view, the objective is to balance the inter-task loss with intra-task loss. This is done automatically by back-propagating multiple task-specific loss values, while regularization is achieved, whereby the

risk of overfitting is lowered. Each task-specific loss function has to be aggregated into a single loss, which the model minimizes. Hence, to maintain stability the losses need to be weighted, usually based on task importance, i.e., weighting for the main task may be different to that for the auxiliary tasks [44]. Another method [45] is to weight by uncertainty, by treating the network as a probabilistic model. The aim is to weight the loss function by maximizing the Gaussian likelihood of the output being the same as the ground truth. Moreover, one can weight by performance, where the authors [46] achieved this by Dynamic Task Prioritization (DTP). DTP employs focal loss [47] as its basis and prioritizes complex tasks by assigning them higher weights, since networks should spend more time learning the difficult tasks.

An example of multi-task learning applied in the medical domain is multi-modality image segmentation [58]. Here the authors trained a single convolutional neural network to segment six tissues in brain MRI, breast MRI, and cardiac CT

angiography images. They were able to achieve comparable performance to models trained individually on each specific task. [59] explored multi-task learning for phenotyping with electronic health records data. Traditionally phenotyping is performed as a supervised learning task; however, the authors in [59] explored when multi-task learning is effective in ascertaining such phenotypes accurately even for rare phenotypes where data availability is low. Their work showed the efficacy of multi-task learning by outperforming models learned only for single tasks.

Neural Architecture Search

This is an extension of “learning to learn.” If one can learn how to best train an AI model, then why not learn to design an AI model itself. This research paradigm focuses on finding optimal methods for designing the best deep learning model architectures. Many state-of-the-art complex deep learning architectures are manually designed by engineers. However, this doesn’t mean every possible configuration has been explored to select the best model option. Ideally a method that could explore different configurations and find the most optimal solution would result in an automatic development of high performance architectures. Neural Architecture Search (NAS) is a sub-field of AutoML (automatic machine learning) and overlaps with meta learning and hyperparameter optimization. The methods in this field of research are usually optimizing the following three categories:

- **Search space:** This defines the network topology. This includes the potential type of operations within the neural network, i.e., convolution, pooling, etc., including how they should be wired to form valid neural networks. Setting and designing this search space requires human expertise and thereby also introduces some level of human biases which may prevent selection of novel architectural building blocks.
- **Search strategy:** This is the algorithm that efficiently explores the search space and finds

the best configuration of models fast while avoiding premature convergence to suboptimal architectures.

- **Evaluation strategy:** The aim is to find models that attain high performance on unseen test data. Hence, it is vital that a metric for the proposed models is measured to determine the quality of the proposed NAS algorithms. Generally, the process of such candidate model evaluation is expensive. Hence, many algorithms innovate by finding methods to save time and computation resources.

The search space in general is quite large, especially if the aim is for the NAS algorithm to start designing a neural network from square one. However, this is possible to do. For example, [48] introduced “NEAT,” a method based on genetic algorithms that optimize for both connections and weights. Like evolutionary algorithms it initiates a population of candidate networks, ranks them, and applies mutation to both weights and connections to the best ranking candidates. This mutation creates a new population and the cycle repeats, thereby selecting the best mutated candidates at the end. However, the total number of possible connection configurations can easily go to infinity, making this system arduous to train. Hence, many methods constrain the search space, by defining a skeleton-like structure [49]. This is typically referred to as cell-based search. Here a user-defined skeleton is developed comprising of multiple cells, as shown in Fig. 4. The search space for NAS thereby becomes limited to that cell, i.e., NAS algorithms must optimize connections and weights for those particular cells as opposed to designing the full connection flow from nothing. This reduces the complexity of search significantly.

Similar to evolutionary algorithms, NAS search strategies are based on sampling from a population of candidate neural networks. The most obvious method is to perform random search as a strategy; however yielding best results can be a hit or miss. [49] utilized reinforcement learning as their search strategy. In particular they define a controller recurrent neural network model which performs classification of the operations to be

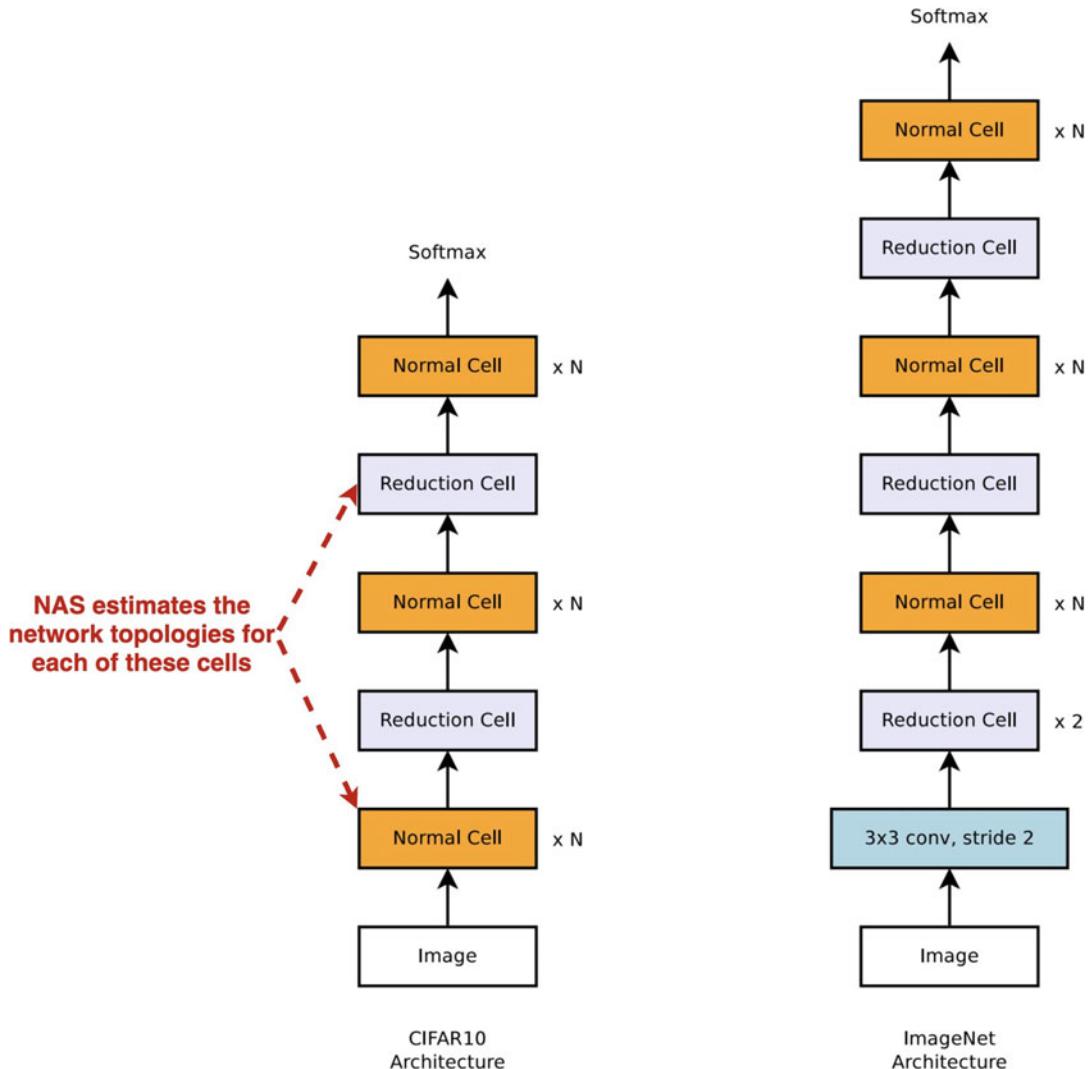


Fig. 4 Displaying cell-based search space for NAS, where the overall information flow is predefined by the user as repeated stacks of cell connected sequentially. NAS

thereby searches for network topologies for these particular cells. (Figure adapted from [49])

selected for building the candidate model (the model to be deployed). As the candidate network is built, it is also trained and evaluated on a given task (i.e., classification). The controller is trained by reinforcement learning where the rewards attained are based on the accuracy level achieved by the candidate network during evaluation. Hence, as the candidate network performs better, the rewards received by the controller are higher and thereby enable the controller to select better operations for the cell.

Furthermore, [50] introduced “AmoebaNet” which further developed on the evolutionary algorithm of “NEAT” as their search strategy. In particular they apply the “tournament selection” method, which at each iteration discards the oldest candidates and favors the younger candidates. These younger candidates at that time may not be the best; however this allows AmoebaNet to explore the search space more, in contrast to NEAT which narrows down the search space to good performing models, sometimes converging

early. Basically they aim to strike a balance between exploration and exploitation. Since designing such models is a gradual process (layer by layer), the complexity increases sequentially. Hence, development should also be sequential such that the complexity can be handled over time. [51] introduced Progressive NAS which uses a similar search space as shown in Fig. 4. Progressive NAS addresses the problem by progressively growing the network, thereby searching for topologies of increasing complexity over time. Instead of reinforcement learning or evolutionary algorithms, the Progressive NAS search strategy is based on model-based Bayesian optimization, similar to the A* search algorithm used widely in robotics.

Lastly, the evaluation strategy is crucial for determining the performance of every possible candidate network, such that in the end the best candidate is selected for inference. One simplest strategy would be to train each candidate model until convergence and measuring the final accuracy. However, this sequential train-converge-evaluate loop would be too slow and computationally expensive. Instead, it is usual to evaluate on a smaller subset of the dataset and train the candidate model for only a small number of epochs. The features learned are enough to evaluate the performance as shown in [49] or to predict the performance itself using time-series regression [52]. Furthermore, [53] achieved an incredible speed-up in performance by conducting parameter sharing among the candidate models. The assumption was that the candidate model topologies can be seen as sub-topologies of a larger conglomerate topology representing the final model. Hence, the candidate model's parameters in these sub-topologies must share the weights for representing the final, enabling convergence to the final model faster.

Conclusion

Currently most AI systems focus on optimizing for single tasks and require huge quantities of data for the learning process. However, this type of AI training procedure is not feasible in the long run, since it is not possible to collect data with respect to

every single potential variation. Especially in the context of medicine, which is a highly dynamic environment, with multiple tasks and ever-changing data. To have a translational AI that can be utilized in the medical industry, it is vital that AI is capable of adapting and continuously learning to adhere to such unseen cases and, more importantly, generalize by retaining its meta-knowledge. As stated, meta learning is still an open question with many new methods being proposed. Although its adoption in medicine may still be limited, it is certainly a direction to move toward. Meta learning is the field that will bring us closer to general intelligence or at least machine learning 2.0. With methods to endow AI with the capabilities of self-designing, self-learning continuously and that too with small data requirements will have an enormous impact in medicine. We must understand that the real world does not comprise of cleanly aggregated balanced data samples; rather it is the opposite, comprising of messy, imbalanced, and highly divergent data. Hence, our AI models should learn to adapt to these situations, instead of forcing us to use collections of data that do not represent the reality, and meta learning is a step toward this achievement.

References

1. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv preprint arXiv:2005.14165. 2020.
2. Hospedales T, Antoniou A, Micaelli P, Storkey A. Meta learning in neural networks: a survey. arXiv e-prints. 2020. arXiv:2004.05439.
3. Zhang XS, Tang F, Dodge HH, Zhou J and Wang F. MetaPred: meta learning for clinical risk prediction with limited patient electronic health records. In: KDD '19: proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019. p. 2487–95.
4. van Sonsbeek T, Cheplygina V. Predicting scores of medical imaging segmentation methods with meta learning. In: Cardoso J, Silva W, Cruz R, Van Nguyen H, Roysam B, Heller N, et al., editors. Interpretable and annotation-efficient learning for medical image computing – 3rd international workshop, iMIMIC 2020, 2nd international workshop, MIL3ID 2020, and 5th international workshop, LABELS 2020, held in conjunction with MICCAI 2020, proceedings. Lecture notes in computer science. Springer; 2020. p. 242–53.

5. Moeskops P, Wolterink JM, van der Velden BHM, Gilhuys KGA, Leiner T, Viergever MA, et al. Deep learning for multi-task medical image segmentation in multiple modalities. In: Ourselin S, Joskowicz L, Sabuncu MR, Unal GB, Wells W, editors. MICCAI (2). Vol. 9901 of lecture notes in computer science. 2016. p. 478–86.
6. Mahajan K, Sharma M, Vig L. Meta-DermDiagnosis: few-shot skin disease identification using meta-learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) workshops. 2020.
7. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. *Nature*. 2015;518(7540): 529–33.
8. Vinyals O, Babuschkin I, Czarnecki MW, Mathieu M, Dudzik A, Chung J, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*. 2019;575(7782):350–4.
9. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: CVPR09. 2009.
10. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-ResNet and the impact of residual connections on learning. In: ICLR 2016 workshop. 2016. <https://arxiv.org/abs/1602.07261>
11. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). 2016. p. 770–8.
12. Demir A, Yilmaz F, Kose O. Early detection of skin cancer using deep learning architectures: Resnet-101 and Inception-V3. In: 2019 medical technologies congress (TIPTEKNO). 2019. p. 1–4.
13. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402–10. <http://jamanetwork.com/journals/jama/fullarticle/2588763>
14. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (Long and short papers). Minneapolis: Association for Computational Linguistics; 2019. p. 4171–86. <https://www.aclweb.org/anthology/N19-1423>
15. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234–40. <https://doi.org/10.1093/bioinformatics/btz682>.
16. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? In: Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger KQ, editors. Advances in neural information processing systems. vol. 27. Curran Associates; 2014. <https://proceedings.neurips.cc/paper/2014/file/375c71349b295fbe2dcda9206f20a06-Paper.pdf>
17. Guo G, Wang H, Bell D, Bi Y, Greer K. KNN model-based approach in classification. In: Meersman R, Tari Z, Schmidt DC, editors. On the move to meaningful internet systems 2003: CoopIS, DOA, and ODBASE. Berlin/Heidelberg: Springer Berlin Heidelberg; 2003. p. 986–96.
18. Hartigan JA, Wong MA. A K-means clustering algorithm. *JSTOR: Appl Stat*. 1979;28(1):100–8.
19. Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in neural information processing systems. vol. 30. Curran Associates; 2017. <https://proceedings.neurips.cc/paper/2017/file/cb8da6767461f2812ae4290eac7cbc42-Paper.pdf>
20. Koch G, Zemel R, Salakhutdinov R. Siamese neural networks for one-shot image recognition. In: Proceedings of the 32nd international conference on machine learning. Lille: Deep Learning Workshop; 2015.
21. Vinyals O, Blundell C, Lillicrap T, Kavukcuoglu K, Wierstra D. Matching networks for one shot learning. In: Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R, editors. Advances in neural information processing systems. vol. 29. Curran Associates; 2016. <https://proceedings.neurips.cc/paper/2016/file/90e1357833654983612fb05e3ec9148c-Paper.pdf>
22. Sung F, Yang Y, Zhang L, Xiang T, Torr PH, Hospedales TM. Learning to compare: relation network for few-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
23. Santoro A, Bartunov S, Botvinick M, Wierstra D, Lillicrap T. Meta learning with memory-augmented neural networks. In: Balcan MF, Weinberger KQ, editors. Proceedings of the 33rd international conference on machine learning. Vol. 48 of proceedings of machine learning research. New York: PMLR; 2016. p. 1842–50. <http://proceedings.mlr.press/v48/santoro16.html>
24. Munkhdalai T, Yu H. Meta networks. In: Precup D, Teh YW, editors. Proceedings of the 34th international conference on machine learning. Vol. 70 of Proceedings of machine learning research. International Convention Centre. Sydney: PMLR; 2017. p. 2554–63. <http://proceedings.mlr.press/v70/munkhdalai17a.html>
25. Ravi S, Larochelle H. Optimization as a model for few-shot learning. In: 5th International conference on learning representations ICLR. 2017.
26. Finn C, Abbeel P, Levine S. Model-agnostic meta learning for fast adaptation of deep networks. In: Precup D, Teh YW, editors. Proceedings of the 34th international conference on machine learning. Vol. 70 of proceedings of machine learning research. International Convention Centre, Sydney: PMLR; 2017. p. 1126–35. <http://proceedings.mlr.press/v70/finn17a.html>
27. Nichol A, Achiam J, Schulman J. On first-order meta learning algorithms. *ArXiv*. 2018. abs/1803.02999.
28. Kirkpatrick J, Pascanu R, Rabinowitz N, Veness J, Desjardins G, Rusu AA, et al. Overcoming catastrophic forgetting in neural networks. 2016. Cite arxiv:1612.00796. <http://arxiv.org/abs/1612.00796>

29. Pierre JM. Incremental lifelong deep learning for autonomous vehicles. In: 2018 21st international conference on intelligent transportation systems (ITSC). 2018. p. 3949–54.
30. Mi F, Lin X, Faltings B. ADER: adaptively distilled exemplar replay towards continual learning for session-based recommendation. In: Fourteenth ACM conference on recommender systems. RecSys '20. New York: Association for Computing Machinery; 2020. p. 408–13. <https://doi.org/10.1145/3383313.3412218>.
31. Rusu A, Rabinowitz C, Desjardins G, Soyer H, Kirkpatrick J, Kavukcuoglu K, Pascanu R, Hadsell R. Progressive neural networks advances in neural information processing systems 29 (NIPS). 2016. abs/1606.04671.
32. Lopez-Paz D, Ranzato MA. Gradient episodic memory for continual learning. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in neural information processing systems, vol. 30. Curran Associates 2017. <https://proceedings.neurips.cc/paper/2017/file/f87522788a2be2d171666752f97dddeb-Paper.pdf>
33. Shin H, Lee JK, Kim J, Kim J. Continual learning with deep generative replay. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in neural information processing systems, vol. 30. Curran Associates; 2017. <https://proceedings.neurips.cc/paper/2017/file/0efbe98067c6c73dba1250d2beaa81f9-Paper.pdf>
34. Zhang Z, Luo P, Loy CC, Tang X. Facial landmark detection by deep multi-task learning. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. Computer vision – ECCV 2014. Cham: Springer International Publishing; 2014. p. 94–108.
35. Teichmann M, Weber M, Zöllner JM, Cipolla R, Urtasun R. MultiNet: Real-time joint semantic reasoning for autonomous driving. In: 2018 IEEE intelligent vehicles symposium, IV 2018, Changshu, Suzhou, China, June 26–30, 2018. IEEE; 2018. p. 1013–20. <https://doi.org/10.1109/IVS.2018.8500504>.
36. Ma J, Zhao Z, Yi X, Chen J, Hong L, Chi EH. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. New York: Association for Computing Machinery; 2018. p. 1930–9. <https://doi.org/10.1145/3219819.3220007>.
37. Dai J, He K, Sun J. Instance-aware semantic segmentation via multi-task network cascades. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). 2016. p. 3150–8.
38. Kokkinos I. UberNet: training a universal convolutional neural network for low-, mid-, and high- level vision using diverse datasets and limited memory. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR). 2017. p. 5454–63.
39. Gao Y, Bai H, Jie Z, Ma J, Jia K, Liu W. MTL-NAS: task-agnostic neural architecture search towards general-purpose multi-task learning. In: IEEE conference on computer vision and pattern recognition (CVPR). 2020.
40. Vandenhende S, Georgoulis S, Gool LV, Brabandere BD. Branched multi-task networks: deciding what layers to share. In: 31st British machine vision conference 2020, BMVC 2020, virtual event, UK, September 7–10, 2020. BMVA Press; 2020. <https://www.bmvc2020-conference.com/assets/papers/0213.pdf>
41. Bruggemann D, Kanakis M, Georgoulis S, Van Gool L. Automated search for resource-efficient branched multi-task networks. In: 31st British machine vision conference 2020, BMVC 2020. 2020.
42. Misra I, Shrivastava A, Gupta A, Hebert M. Cross-stitch networks for multi-task learning. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). 2016. p. 3994–4003.
43. Gao Y, Ma J, Zhao M, Liu W, Yuille AL. NDDR-CNN: layerwise feature fusing in multi-task CNNs by neural discriminative dimensionality reduction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). 2019.
44. Xu Y, Liu X, Shen Y, Liu J, Gao J. Multi-task learning with sample re-weighting for machine reading comprehension. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (Long and short papers). Minneapolis: Association for Computational Linguistics; 2019. p. 2644–55. <https://www.aclweb.org/anthology/N19-1271>
45. Cipolla R, Gal Y, Kendall A. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: 2018 IEEE/CVF conference on computer vision and pattern recognition. 2018. p. 7482–91.
46. Guo M, Haque A, Huang DA, Yeung S, Fei-Fei L. Dynamic task prioritization for multitask learning. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, editors. Computer vision – ECCV 2018. Cham: Springer International Publishing; 2018. p. 282–99.
47. Lin T, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell. 2020;42(2):318–27.
48. Stanley KO, Miikkulainen R. Evolving neural networks through augmenting topologies. Evol Comput. 2002;10(2):99–127.
49. Zoph B, Vasudevan V, Shlens J, Le QV. Learning transferable architectures for scalable image recognition. In: 2018 IEEE/CVF conference on computer vision and pattern recognition. 2018. p. 8697–710.
50. Real E, Aggarwal A, Huang Y, Le QV. Regularized evolution for image classifier architecture search. Proc AAAI Conf Artif Intell. 2019;33(01):4780–9. <https://ojs.aaai.org/index.php/AAAI/article/view/4405>
51. Liu C, Zoph B, Neumann M, Shlens J, Hua W, Li LJ, et al. Progressive neural architecture search. In: Proceedings of the European Conference on Computer Vision (ECCV). 2018.
52. Baker B, Gupta O, Raskar R, Naik N. Accelerating neural architecture search using performance

- prediction. International conference on learning representations. 2018.
53. Pham H, Guan M, Zoph B, Le Q, Dean J. Efficient neural architecture search via parameters sharing. In: Dy J, Krause A, editors. Proceedings of the 35th international conference on machine learning. Vol. 80 of proceedings of machine learning research. Stockholm: PMLR; 2018. p. 4095–104. <http://proceedings.mlr.press/v80/pham18a.html>
54. Ruder S, Bingel J, Augenstein I, SØgaard A. Sluice networks: learning what to share between loosely related tasks. arXiv: 2017, abs/1705.08142.
55. Guha Roy A, Siddiqui S, Pölsterl S, Navab N, Wachinger C. ‘Squeeze & excite’ guided few-shot segmentation of volumetric images. Med Image Anal. 2020;59:101587. <https://doi.org/10.1016/j.media.2019.101587>.
56. Prabhu V, Kannan A, Ravuri M, Chaplain M, Sontag D, Amatriain X. Few-shot learning for dermatological disease diagnosis. In: Proceedings of the 4th machine learning for healthcare conference. PMLR; 2019. p. 532–52.
57. Altae-Tran H, Ramsundar B, Pappu A, Pande V. Low data drug discovery with one-shot learning. ACS Cent Sci. 2017;3(4):283–93.
58. Moeskops P, Wolterink J, Velden B, Gilhuijs K, Leiner T, Viergever M, Isgum I. Deep learning for multi-task medical image segmentation in multiple modalities. In: MICCAI 2016, LNCS vol. 9001 Part 2. 2016. p. 478–86.
59. Ding DY, Simpson C, Pfahl S, Kale DC, Jung K, Shah NH. The effectiveness of multitask learning for phenotyping with electronic health records data. Pac Symp Biocomput. 2019;24:18–29. PMID: 30864307; PMCID: PMC6662921.



Artificial Intelligence in Medicine Using Quantum Computing in the Future of Healthcare

30

Joseph Davids, Niklas Lidströmer, and Hutan Ashrafian

Contents

Introduction	424
History	426
The Basics of Quantum Computingand Quantum Machine Learning	426
Types of Quantum Computers	426
Companies	429
Basic Anatomy of a Quantum Versus Classical Computing Circuit	429
The Bit and the Classical Logic Gate	429
The Qubit and the Quantum Gate	431
State Vectors	432
The Quantum Gates	433

J. Davids (✉)
Imperial College London NHS Trust, London, UK

Hamlyn Centre for Robotics and Artificial Intelligence,
Department of Surgery and Cancer, Imperial College
London, London, UK

National Hospital for Neurology and Neurosurgery Queen
Square, London, UK
e-mail: j.davids@imperial.ac.uk;
meniscusmedical@gmail.com

N. Lidströmer
Department of Women's and Children's Health, Karolinska
Institutet, Stockholm, Sweden
e-mail: niklas.lidstromer@ki.se; niklas@lidstromer.com

H. Ashrafian
Department of Surgery and Cancer, Imperial College
London NHS Trust, London, UK

Institute of Global Health Innovation, Imperial College
London, London, UK

Hamlyn Centre for Robotics and Artificial Intelligence,
Department of Surgery and Cancer, Imperial College
London, London, UK
e-mail: h.ashrafian@imperial.ac.uk

Quantum Superposition and Machine Learning Applications	433
Theoretical Medical Applications	433
Quantum Tunnelling and Machine Learning Applications	435
Quantum Entanglement and MachineLearning Applications for Medicine and Surgery	436
Basic Mathematical Formalism of Entanglement Applied to Medicine	436
Important Quantum Computing Algorithms and Quantum Machine Learning Algorithms	438
Quantum Fourier Transform	439
Shor's Algorithm	439
Grover's Search Algorithm	439
Deutsch-Jozsa Algorithm	440
Bernstein-Vazirani Algorithm	440
Quantum Hidden Markov Chain Algorithm	440
Quantum Natural Language Processing	440
Quantum Annealing and Quantum Neural Networks	440
Quantum Enhanced Reinforcement Learning	441
Quantum Phenomenon in Disease	441
Quantum Computing in Healthcare	441
Future Translational Considerations for E-Healthcare	442
Ethico-Legal Implications and Considerations	442
Summary Remarks	443
References	443

Abstract

The concept of quantum computing has evolved over nearly a century to a point now where it is no longer science-fiction. However, conceptual extensions of quantum computation and many body systems to quantum clinical medicine and quantum surgery are completely original areas that are yet to be realized in terms of their development and full potential. Novel formalisms and approaches will have to evolve to enable these areas to fully materialize and mature into safe clinical applications that will benefit mankind.

Nevertheless, factors paving the way for this exciting area of medical and future surgical science include the exponential advances in computational power gained through newly evolved mathematical formalisms for algorithmic design such as quantum mechanics, category theory, quantum algebraic geometry and others, coupled with advances in precision nanoengineering.

This chapter offers a cursory non-exhaustive primer to the topic of quantum

machine learning for medicine, surgery and healthcare, highlighting some of the areas where the authors theorise that quantum computing will help augment medicine, surgery and healthcare to usher in next-level precision medical diagnostics and therapeutics. In the not-too-distant future, quantum medicine and surgery will offer the ability to re-calibrate the continuous state of flux that occurs in conditions like cancer and neurological diseases to a manageably consistent curative state.

Introduction

For centuries we have sought to harness analytical techniques to understand scientific processes including how the human body functions on a molecular level, and we have been rewarded by identifying key behaviors of what constitutes matter on the atomic scale. We now appreciate the discovery of several elementary particles within the universe ranging from photons, protons, electrons, neutrons, neutrinos, muons, tau, hyperons,

positrons, etc.; most of which are being or have already been harnessed for medical applications [1]. With the limited tools we have at our disposal, we have gained some understanding of how the universe and its myriad of constituents can be modelled to follow a very rudimentary and consistent pattern of atomic and molecular architecture. The Niels Bohr model of the basic atomic structure is very familiar to both non-scientist and scientists alike. We now accept that the atom has a nucleus with fundamental particles like electrons arranged in their probabilistically predictable orbitals in a quantized precise energy domain. This led to discoveries by many scientists like Erwin Schrödinger, whose wavefunction equation laid the foundations for the advancement of quantum mechanics. This was further explored by Nobel Laureates like Richard Feynman who gave the world a detailed exposition of quantum mechanics through his famous lectures [2]. Louis de Broglie described the wave-particle duality of matter and helped us to fundamentally appreciate quantum behavior in all matter.

The collective flow of electrons modelled as charged particles enables electricity and electromagnetism, but the movement of electrons could include their spin with respect to their axis or in relation to other electrons and atoms. Whether they are paired or not and how they interchange between energy-dictated quantized levels has provided added insights into what permits quantum optical phenomena, such as fluorescence and several others, to be respected. Many of these are now used in medical diagnostics. Unsurprisingly the spin of the nucleus itself also allows unique properties to be appreciated and observed, which are necessary for quantum computing applications in the field of spintronics.

Newer discoveries from using the large hadron and large electron-positron colliders continue to surface including the identity of stable and unstable matter particles at a subatomic level, classified into respective quarks, leptons (collectively known as fermions), and bosons [1]. However, it is generally agreed that in stable terrestrial form, matter is fundamentally composed of atoms with electron orbitals that can also generate quantized electromagnetic fields in a constant state of flux. This is an area we have implemented to develop medical

imaging technologies like nuclear magnetic resonance imaging, which leads to exploiting magnetic field alignments of the hydrogen atoms that forms the make-up of the human body. Our earthly chemical make-up remains mostly either organic hydrocarbon based as life-forms or inorganic with elements such as silicon being abundant in various unaltered states. Silica and other inorganics have allowed the development of semi-conductors for which the transistor technology underscores the classical computer on which we depend.

Through transistors and other tools arranged in a very precise and specific way, and for which we now have replicable blueprints, we have designed algorithms that control the flow of electrons within transistors in a very precise way. This is one fundamental view of how machine learning works where high-level programs are written in syntactically sound programming languages like Python and C++, either respectively interpreted or compiled to machine code using assemblers, but in general explains how software communicates with the transistors in computing. The behavior of some of these transistors can be explained at a fundamental quantum level in some form using quantum mechanics and include principles of quantum interference, quantum entanglement, quantum tunnelling and quantum superposition.

Carbon-based systems like graphene and diamonds can also behave as quantum computers to facilitate quantum information transfer and sensing at a fundamental level with potential for other hydrocarbon-based systems like polycyclic aromatic hydrocarbons being demonstrated [3–6]. Through advances in imaging technology, we have been able to appreciate and harness these molecular interactions into the development of techniques that allow us to even reconstruct matter from thin air using carbon capture as an example of modern chemistry [7], precisely control the flow of electrons in Bose-Einstein condensates [8], precisely move atoms from one location to another and several other feats that two centuries ago would have been deemed impossible.

Another area where we have exploited a quantum system to discover other systems with quantum potential involved leveraging X-ray diffractometry, which aided the discovery of the deoxyribonucleic acid in the 1950's- a molecular

system that facilitates the blueprint for the construction of the organic being in its entirety for which its counterpart in the inorganic domain is yet to be unveiled [9]. DNA technology has now led to the ability to clone Dolly the sheep and other organisms, but it is the quantum information associated with it that is becoming a growing area of interest [10–12]. Suffice to say that this has interestingly allowed the appreciation of quantum molecular interactions and information flow through dynamic signal processing and precision recordings that are only recently becoming accessible through quantum computing. The potential for quantum biology using DNA nanotechnologies like DNA origami [13], DNA for quantum information processing and personalized cryptographic encoding for healthcare data using DNA makes this an even more attractive area of recent research focus [14].

As a consequence of the aforementioned, one could model the human body as an entropic quantum information system where Shannon's information theory may apply on a quantum level and become necessary for exploration [15, 16]. Clearly, there is a need to at least consider dispelling the myths associated with the fact that the human body has no quantum potential- a topic of considerable ongoing debate as very few people know about quantum physics.

History

Our generous universe comes equipped with the ability to compute.

David Bacon

The ability to adopt and effect computational superiority for problem solving has been with us well before Charles Babbage's "difference engine" concept sparked our curiosity about computers in 1833 [17]. However, the concept of quantum computation has been hidden in plain sight for nearly two centuries. The above quote from David Bacon sums it up best. The fifth Solvay conference on physics was held in October 1927 and is thought to be the birthplace of quantum theory with over 17 Nobel Laureates in

attendance arguing for and against various concepts of photon and electron behavior. We summarize the history of quantum computing in Fig. 1, looking at some of the modern landmark developments since the 1930's that propelled the field forwards towards its physical realizations in the mid-late twentieth century until quantum supremacy was attained in 2019/2020 [18–20]. Many computational theories evolved since the 1930's. The Church-Turing thesis in 1939, laid important foundations about the nature of computable functions in an attempt to resolve the Hilbert-Ackermann "Entscheidungsproblem" ("decision problem" to use mechanical functions to separate mathematical truths from falsehoods) [21–23]. This significantly contributed to developing the Turing machine and the abstract quantum variants of which became heavily researched and initially formulated by David Deutsch allowing for any quantum algorithm to be expressed as a particular Turing machine [22]. Another extension of the Church-Turing theory called the feasibility theory was presented for Quantum Complexity theory and is a foundation for the Bernstein-Vazirani quantum algorithm presented later [24, 25].

The Basics of Quantum Computing and Quantum Machine Learning

Types of Quantum Computers

Despite the digital technological revolution in classical computing steering us into a wave of spectacular scientific and medical discoveries, limitations have also left us with additional challenges needing advanced computational improvements to overcome [11]. These limitations of silicon-based integrated chips mean newer approaches need to be considered for computation. Quantum computers boast superior computation over their classical equivalents in a 1 million to 1 operational ratio and are intended to be useful for problems that classical computers would be unable to solve [11]. Quantum supremacy have been achieved both through semiconducting qubit computers and optical/photonics-based quantum

computers [18–20, 26]. A non-exhaustive list of their current forms with some in physical realizations is provided below:

1. Superconducting Quantum Computers [20],
2. Ion-trapping-based Quantum Computers [27],
3. Quantum Annealing & Adiabatic Quantum Computers [28–30],
4. Quantum Computing Using Engineered Quantum Wells [31],
5. Quantum Computing Using Neutral Atoms [32],
6. Anyon-based Topological Quantum Computers [33],
7. Linear Optical Quantum Computers [34],
8. Transistor-based Quantum Computers [35],
9. Coupled Quantum Wire-based Quantum Computers [36],
10. Optical Lattice System-based Quantum Computers [37]
11. Quantum Dot Quantum Computers [38],
12. Nuclear Magnetic Resonance (NMR) Quantum Computing [39],
13. Kane (Hybrid quantum Dot and NMR) Quantum Computers [40],
14. Electrons On Helium Quantum Computers [41],
15. Inorganic Crystal-based Quantum Computers [42],
16. Cavity Quantum Electrodynamical Quantum Computers [43],
17. Carbon Nanosphere-based Quantum Computers [44],
18. Fullerene-based Quantum Computers [45],
19. Diamond-based Quantum Computers [6, 46]

A)

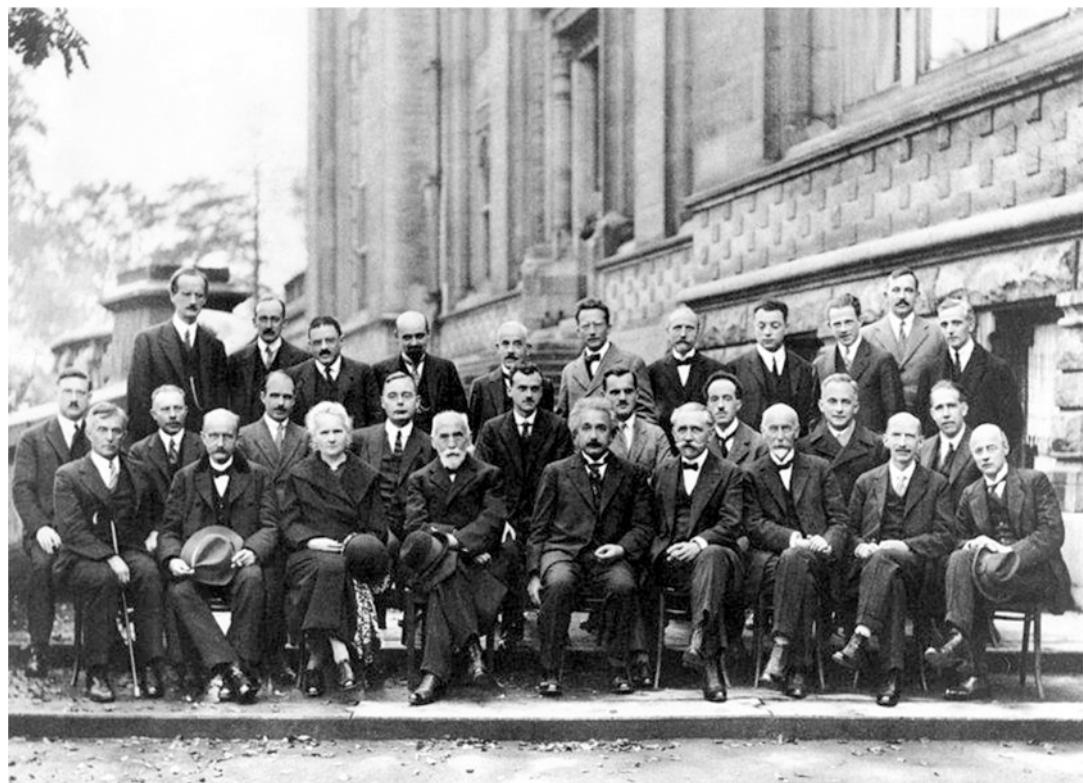


Fig. 1 (continued)

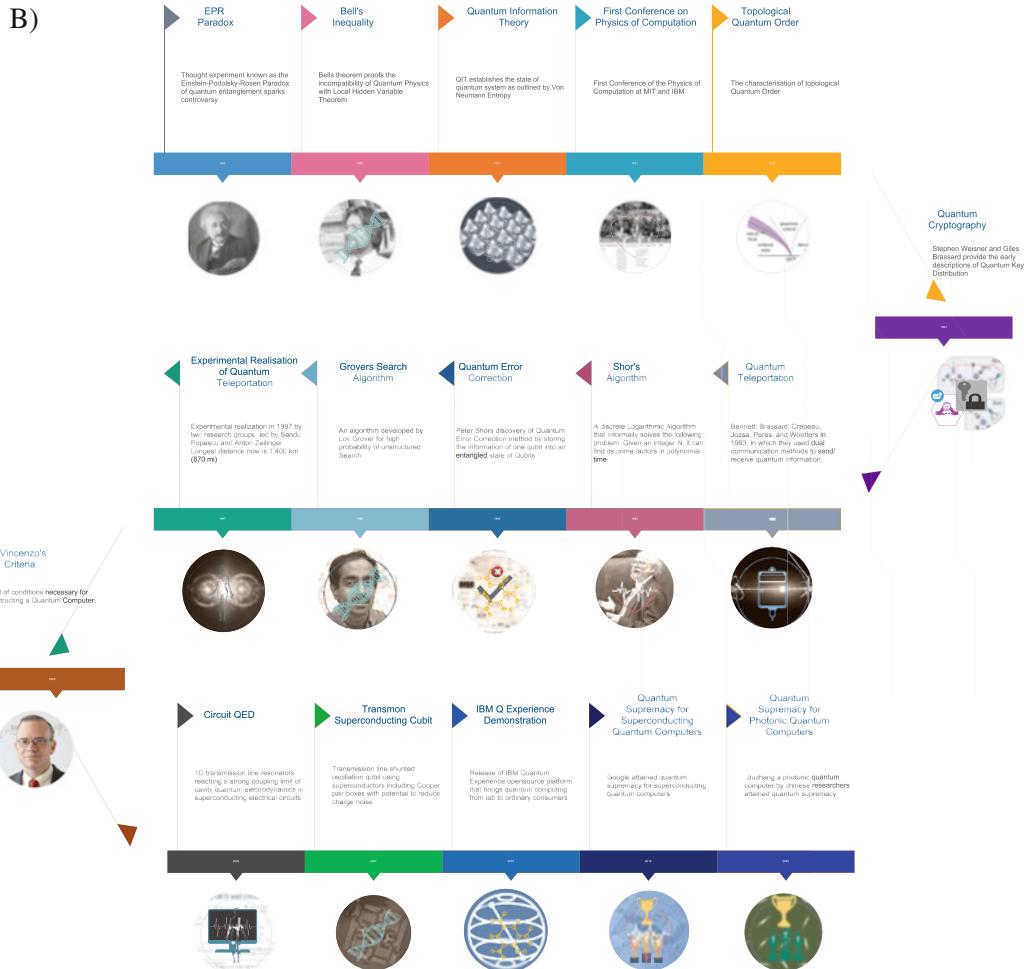


Fig. 1 History of Quantum Computing. Top Picture (a) is from the fifth Solvay Conference that started the discussion about quantum theory focusing on electron and photons. (Source https://en.wikipedia.org/wiki/File:Solvay_conference_1927.jpg). Bottom Picture (b) illustrates how modern

quantum computing has evolved over the years with some key landmarks; time-line was created using EdrawMax software, but for detailed look at some of the milestones see https://en.wikipedia.org/wiki/Timeline_of_quantum_computing_and_communication

20. Bose Einstein Condensate Quantum Computers [47]

The above operate on four main currently applicable computational models with a fifth being the Quantum Turing Machine yet to reach full physical realization. The current four models are:

- Adiabatic quantum computing
- One-way quantum computing
- Gate array quantum computing
- Topological lattice-based quantum computing

Current quantum computers have a significant degree of quantum noise related error and as such they are known as noisy intermediate-scale quantum computers (NISQ) leading to decoherence. This means only a limited number of operations can be computed on a qubit before it loses its quantum state leading to random noise as the output. Figures 2 and 3 illustrate the current quantum computer and chip. Newer computational models will evolve for quantum platforms that will be useful for future medical applications.



Fig. 2 Image of a Quantum Computer demonstrating that most of the spaces have to be achieved using a dilution refrigerator cooling to around 10–20 millikelvin greater than the temperature of outer space. With specialised cables that transmit microwave pulses to the chip shown in Fig. 3 An IBM Quantum Computer; Image Source: <https://www.research.ibm.com/ibm-q/network/> and [64]

Companies

Some of the companies developing quantum computing include, *Google's Sycamore, Atom Computing (a quantum computing hardware company specializing in neutral atom quantum computers), Xanadu, IBM (IBM Q), ColdQuanta, Zapata*

Computing, Azure Quantum, D-Wave, Strangeworks [22, 48–54]. Shenzhen Spin-Q Technology is a Chinese founded company aiming to release the first 2-qubit desktop quantum computer for schools [55].

This section aims to guide the reader towards areas of specific enquiry, investigation, knowledge discovery and as such not meant to be exhaustive. Its aim is to introduce the medical reader to some of the notations that they will need to be familiar with when reading papers going forward. The authors however do highlight a few new ideas underpinned by fundamental quantum principles of superposition, entanglement and quantum tunnelling as the field is still very new. For all these areas, we will discuss the potential impacts and where in-roads can be made for medicine and surgery.

Basic Anatomy of a Quantum Versus Classical Computing Circuit

This section discusses how quantum circuitry differs from their classical counterparts and how this applies to the design of newer medical devices that one would aim to use for efficient quantum computations. We follow this with an explanation of the quantum phenomena that makes computing superior to its classical equivalent.

The Bit and the Classical Logic Gate

Classical computation occurs with operations and information encoded in a binary format that arises through a combination of bits of information. According to Shannon's information theory, a binary digit averaged forms one bit of Shannon information. The bits can be isolated states of either 0 (False) or 1 (True) in a Boolean logic state analogously behaving as a gate that opens and closes to allow or restrict information flow. This bit of information is the binary classification system that underscores how the transistor-based logic gates behave. These Boolean logic gates, in classical computers enable operations like AND, OR, XOR, NOT, NAND, NOR, and XNOR to be

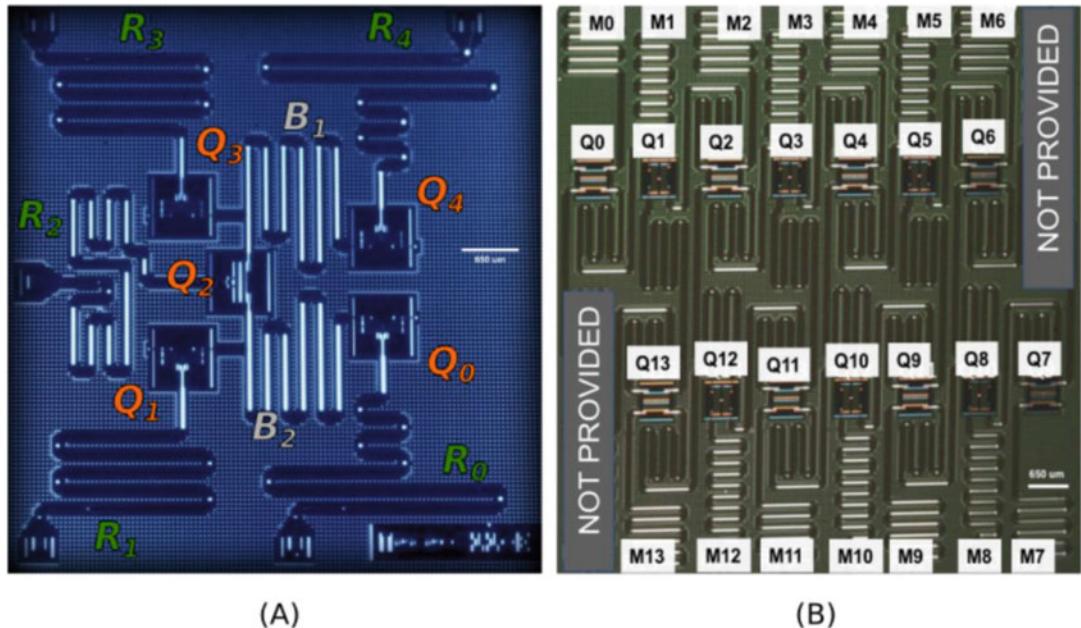


Fig. 3 Quantum Chip photos and layout example courtesy of IBM open-source quantum. Melbourne **(b)** and Tenerife chipsets **(a)**, the Chip layout of a 5-qubit (ibmqx4) and 14-qubit (IBM Q 14 Melbourne) quantum

achieved where the resulting value of 0 and 1 from the combination of bits in this way executes the equivalent of a True or False to either propagate or restrict information flow respectively (see Fig. 4). These gating mechanisms also underly how computers can understand higher level human languages converted into their lower-level equivalents to aid the execution of algorithms by the hardware. The combinatorial mixture of these gates is how operations are carried out in a functioning computer. Additionally, the combination of these bits of information in a specific well-defined order allows intelligent information processing and transfer. This phenomenon of True/False gated signaling also routinely happens in biology. Common examples include the conduction of action potentials and neuronal depolarization which is a gated threshold-activated, ion-channel system operating on an all-or-nothing basis. Here, subthreshold excitatory post-synaptic vesicular release of neurotransmitter stimulation of the post-synaptic membrane does not facilitate and propagate an action potential; the on-signal/True.

Classical computing techniques have been harnessed in modern medical and surgical devices

processors [51]. The chipset is sensitive to radiofrequency, noise, temperature, etc. and require cooling as discussed in Fig. 2

that compute diagnostic information, and require additional methods of convertible analogue-to-digital bitwise sensing and post-processing. This feat is also necessary for designing robust encryption/decryption systems. Pitfalls of classical computation involve the impaired ability to rapidly factorize large prime numbers. Something that the healthcare and other sectors require to safeguard clinical information. Prime number factorization function is one that also models exponential growth and is something that our current classical computers struggle to overcome. The security of healthcare and surgical tele-robotic systems as well as medical diagnostic and therapeutic devices, which all depend on robust public and private key cryptographic encryption systems obey the prime factorization archetype. Thus, as quantum computing evolves, these encryption systems can be put at risk and need more sophisticated computational methods.

Another area worth considering is our current suboptimal infrastructure to quickly model pandemics which follow the exponential growth and decay equation. This field could also benefit from the efficient speedups that classical computation

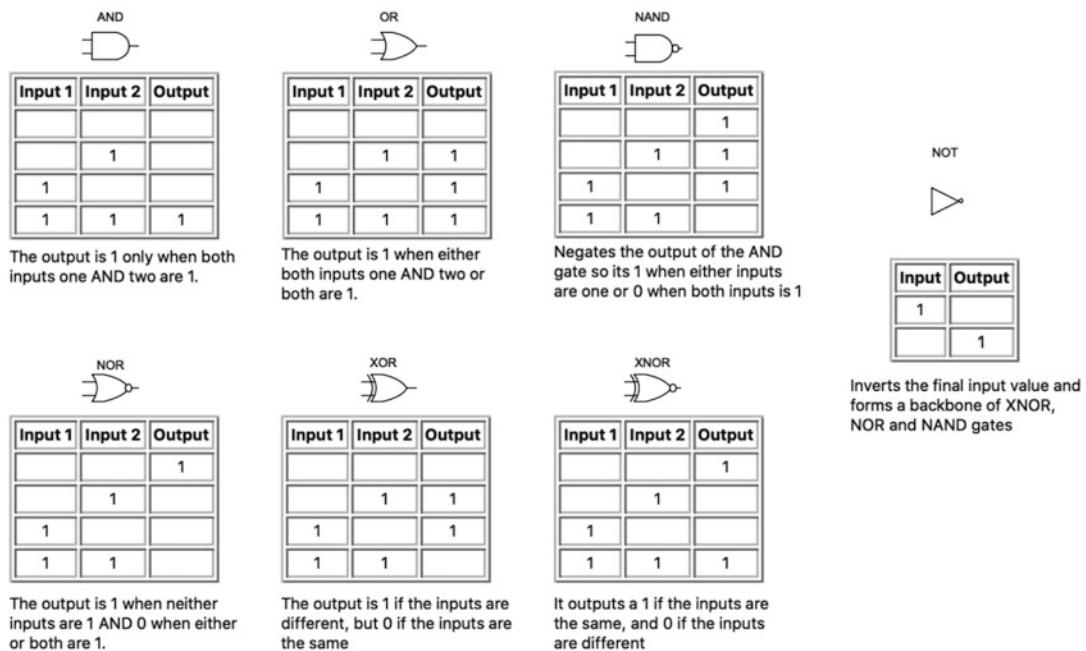


Fig. 4 Illustrates classical logic gates in digital electronics and their truth tables of computational outputs. Specific combinations of these Boolean logic gates enable all

classical digital computation. Contrast this to Fig. 7, which demonstrates some of the quantum logic gates

currently lacks. This together with impaired and suboptimal approaches in the probabilistic resolution of simulating molecular bonding interaction information in disease warrants other more advanced computational alternatives far superior to any classical purview of exascale supercomputing [56, 57].

The Qubit and the Quantum Gate [58]

On the contrary, quantum logic gates are unitary matrices, which operate on quantum bits (qubits) of information. Qubits allow the phenomenon of what is known as coherent superposition of states to occur. A single qubit has a vector representation whereby in addition to the presence of single independent Boolean binary-style logic states in either a $|0\rangle$ or $|1\rangle$ (pronounced Ket 0 or Ket 1) or spin state of an up $|\uparrow\rangle$ or down $|\downarrow\rangle$, a superposition represented as $|\psi\rangle$ can also exist. The medical reader is also directed to the institute of advanced study for further simplification [59].

These two states can be available simultaneously giving either an $\alpha|0\rangle + \beta|1\rangle$ or $\alpha|\uparrow\rangle$

$+ \beta|\downarrow\rangle$ configuration. Alpha (α) and Beta (β) are usually complex probability amplitudes, hence squaring the magnitudes $|\alpha|^2$ or $|\beta|^2$ gives the actual probability that its corresponding state of 0 or 1 will occur. The sum of these two probability amplitudes must be equal to 1. Also, these probability amplitudes can be complex vectors. The significance is the thorough capacity for the quantum system to holistically parallelize resources for information inquisition and rapidly explore all probable permutations of information flow simultaneously. The underlying scientific ideas are based on Heisenberg's uncertainty principle and paradigms like the De Broglie wave-particle duality in atomic scale quantum systems [60, 61].

Figure 5 illustrates this concept well using a simple maze analogy. It is only when one chooses to measure which observable state the quantum system is in that it collapses into a single measurable state, otherwise it remains in superposition of multiple states, similar to the paradigm of Schrödinger's cat, which is simultaneously both dead and alive until it is observed. This collapse of the quantum state

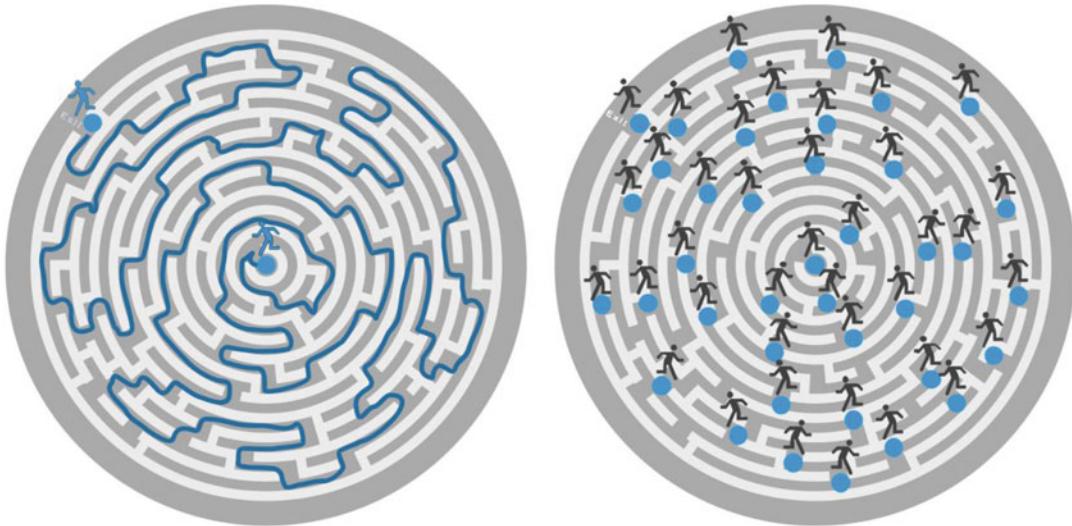


Fig. 5 Maze-like concept for experiencing how a quantum system behaves versus the classical equivalent. The quantum maze phenomenon has been observed in photonic quantum systems in biology [98]. (a) Is a classical system maze. The blue classical particle without any quantum ability brute forces it's way and explore every sequential eventuality before it finds its way out. At each level of the maze, it will attempt to seek a True or False answer to the question of whether an opening exist to move to the next level of the maze. (b) Represents a quantum system maze.

The blue quantum particle searching for the quickest exit path is able to explore all probabilistic eventualities using the principles of (1) superposition exploring eventualities simultaneously, (2) principles of entanglement where the fundamental particle behaves as if it has multiple copies of itself and omnipresent to explore all possible paths and areas of the maze, and (3) principles of quantum tunnelling (where it has the ability to pass through obstacle or a boundary in its path to the other side of the maze). This figure was created with EdrawMax Software

occurs when the system has been perturbed by measurement. The significance here is that with quantum computation, information classification can be very rapid. Decryption of encrypted systems based on pre-quantum public and private cryptosystems takes a fraction of the time it would otherwise take for a classical computer to break. It is therefore prudent to design algorithms for post-quantum cryptographic encryption to mitigate against security vulnerabilities inherent in electronic devices and database information management systems used in medicine, surgery and healthcare.

Mathematically, the above concepts of the various states can be encoded as state vectors using the dynamical representations/pictures of quantum mechanics, which mainly consist of Schrödinger's and Heisenberg's picture with an intermediate/Dirac picture [62]. Various other mathematical formulations exist with as many as nine explored in reference [63].

State Vectors

To appreciate the function of the quantum logic gates one must first understand the quantum state.

In quantum mechanics the state is an entity that provides a mathematical probability distribution of where the fundamental quantum particle (electron, photon etc.) is at a particular point in time and space. This is necessary for us to obtain a possible outcome when we measure it. The measurement of the particle or molecule can be its momentum, spin, color change, interference pattern, bond length, force interaction, thermodynamics through Brownian motion, entropy, etc. Knowledge of the state and the rules of the quantum systems appears sufficient to predict the behavior of the system. Additionally, the quantum states can be mixed to give new states called mixed quantum states, otherwise they are referred to as pure quantum states. Pure quantum states are represented by a projective

ray over a Hilbert space vector over complex numbers. The space that describes a qubit can be explained using this complex projective ray and this projection can be represented on a specialist sphere known as the Bloch sphere [64]. Figure 6 illustrates the Bloch sphere [64].

The quantum state of a qubit is therefore a linear superposition of its orthonormal basis vectors and it is written in Dirac's Bra ($\langle \cdot |$) Ket ($| \cdot \rangle$) notation using angle brackets. The Ket is a linear vector in the Hilbert's space denoted as $|0\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ or $|1\rangle = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

A two-qubit system can be represented by:

$$|xy\rangle = |x\rangle \otimes |y\rangle = \mathbf{r}_{00}|00\rangle + \mathbf{r}_{01}|01\rangle + \mathbf{r}_{10}|10\rangle + \mathbf{r}_{11}|11\rangle \rightarrow \begin{bmatrix} \mathbf{r}_{00} \\ \mathbf{r}_{01} \\ \mathbf{r}_{10} \\ \mathbf{r}_{11} \end{bmatrix}$$

As well as single-qubit systems, there can be other multi-qubit systems like the above where the tensor/Kronecker product \otimes combines quantum states. Such multi state systems have representational circuits which can undergo the phenomenon of quantum entanglement discussed later.

The Quantum Gates [58]

The quantum gate is mathematically represented as a unitary operation and acts on the qubits, which are in various states as described by the example vectors above. Figure 7 illustrates what a quantum gate looks like. Generally, a gate can act on n qubits, where n is the number of qubits represented by a $2^n \times 2^n$ matrix. How the gate behaves on a specific quantum state can be derived by multiplying the Ket state vector (e.g., $|\psi_A\rangle$) by the gate's unitary matrix M_U

$M_U|\psi_A\rangle = |\psi_B\rangle$. Such an operation results in a new quantum state vector being created.

An example is the Hadamard gate, which acts on a single qubit to equalize the probability of both states 0 or 1 occurring and hence it is the

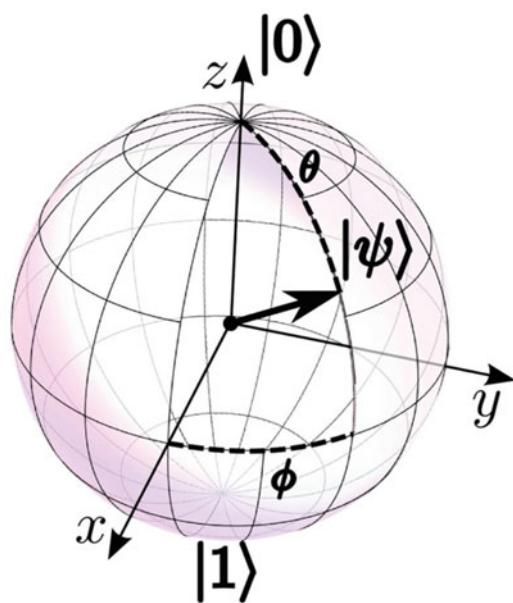


Fig. 6 A computational quantum particle – qubit – represented as a unit (Bloch) sphere [64]. Unlike regular bit that takes values of 0 or 1, a qubit can take any values on a sphere, as represented by two angles. The poles correspond to classical 0/1 bit values. (Image source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6205278/figure/f2-ms115_p0463/)

gate to create a superposition of states. The Hadamard gate uses a special unitary matrix called the Hadamard matrix shown in Fig. 7 and represents a π rotation on the block sphere about the axes $(\hat{x} + \hat{z})/\sqrt{2}$.

Several gates exist and the common ones are illustrated in Fig. 7. The Pauli X gate is equivalent to the classical NOT gate. Squaring the Pauli gates results in an identity matrix. Others include the Toffoli, Fredkin, CCNOT, SWAP, 2-qubit Ising coupling gates (native to trapped ion channel quantum computers), 3-qubit Deutsch gates, etc.

Quantum Superposition and Machine Learning Applications

Theoretical Medical Applications

At a fundamental level, applying the above principles of gate-linked qubit quantum superposition to machine learning and clinical practice

Operator	Gate(s)	Matrix
Pauli-X (X)		$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$
Pauli-Y (Y)		$\begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}$
Pauli-Z (Z)		$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$
Hadamard (H)		$\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$
Phase (S, P)		$\begin{bmatrix} 1 & 0 \\ 0 & i \end{bmatrix}$
$\pi/8$ (T)		$\begin{bmatrix} 1 & 0 \\ 0 & e^{i\pi/4} \end{bmatrix}$
Controlled Not (CNOT, CX)		$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$
Controlled Z (CZ)		$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$
SWAP		$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$
Toffoli (CCNOT, CCX, TOFF)		$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$

Fig. 7 Some very common quantum gate circuitry like the Hadamard and Pauli Gates. Contrast these unique gates to the classical logic gates in Fig. 4. (Image Source https://upload.wikimedia.org/wikipedia/commons/e/e0/Quantum_Logic_Gates.png)

would see the potential for gate discoveries acting as operators that may potentially lead to the modelling of various diseased states theoretically simplified here with a mathematical generalization:

$$M_U \text{ operator } |\psi_{\text{Norm Tissue}}\rangle = |\psi_{D \text{ State A Early}}\rangle + |\psi_{D \text{ State B Intermediate}}\rangle + |\psi_{D \text{ State C Final}}\rangle + |\psi_{D \text{ State error}}\rangle$$

The operator M_U would be a quantum disease gating matrix that acts in some form on the normal tissue state matrix to lead to an altered superposition of diseased states. This will be feasible through clever design of quantum machine learning powering operator-dependent surgical tools, devices, and treatment platforms including brain machine interfaces.

A most recent application leveraging quantum mechanical behavior and predictive machine

learning in the brain is seen in Elon Musk's Neuralink platform, which draws from quantum mechanical properties of fabricated nanomaterials for sensing and information processing [65].

Quantum Machine learning algorithms derived from newer gating combinations will also enable the potential to better delineate these intermediate disease states that lead to rapid progression into advanced forms to either prevent progression, slow down progression or potentially reverse disease.

Quantum Tunnelling and Machine Learning Applications

Quantum tunnelling is a phenomenon whereby a wavefunction (wave-particle duality paradigm) propagates through a barrier or obstacle as if the obstacle was not present. Figure 8 illustrates this phenomenon well.

Derivations of tunnelling arise from both the Helmholtz equation and Schrödinger's time-dependent equation that unites the Galilei and Lorentz invariants, as well as Einstein's energy relations and this allows one to calculate the wave number of a fundamental particle with a degree of precision. Evanescent tunnelling arises when the Helmholtz equation loses its Lorentzian invariant [61]. The boundary interaction time of the wavefunction does not depend on the length of the barrier. Table 1 demonstrates the tunnelling times for some of the common fundamental particles in quantum computing. Dynamic and Chaos-assisted tunnelling are two forms of this phenomenon that enables machine learning methodologies on a quantum scale.

This property and the speed-up that it offers makes this attractive to the scientific and medical community when it comes to building machine learning algorithms for medicine that can be

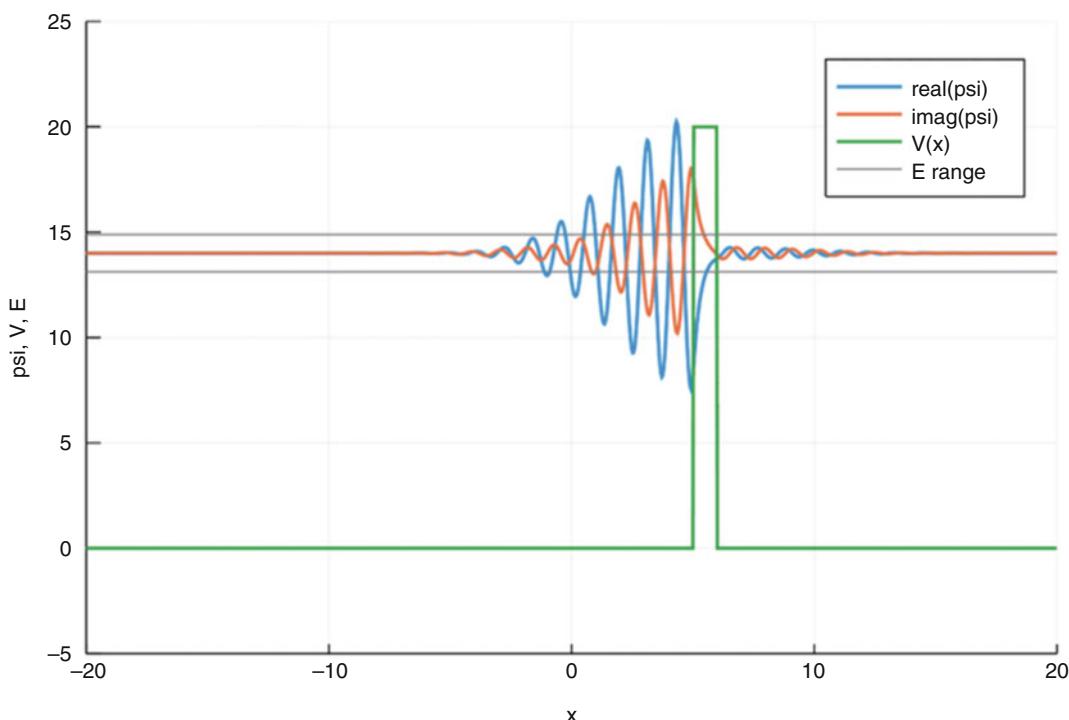


Fig. 8 Illustrates the quantum tunnelling phenomenon. A wave can propagate through a barrier of a particular size to the other side of the barrier. (Source BeCarlson <https://upload.wikimedia.org/wikipedia/commons/4/48/E14-V20-B1.gif>)

Table 1 Demonstrates the tunnelling times for some of the common fundamental particles in quantum computing. It is clear that comparable speed-ups are seen for photonic methods over electron methods

Tunnelling barriers	Tunnelling Time τ	Transmission Time $T = 1/v$	References
Photonic lattice	Between 2.13 femtoseconds (10^{-15}) and 2.7 femtoseconds	Between 2.34 femtoseconds and 2.7 femtoseconds	[61, 66, 67]
electron field-emission tunnelling	7 femtoseconds	6 femtoseconds	[68]
electron ionization tunnelling	≤ 6 attoseconds (10^{-18})	Not reported	[69]
Acoustic (phonon) tunnelling	Between 0.8 microseconds (μs) and 0.9 milliseconds (ms)	Between 1 μs and 1 ms	[70, 71]

implemented and these quantum algorithms are superior to classical algorithms.

Scanning Tunnelling Microscopy has been used in medical imaging to image disease processes like Amyloid protein deposition in Alzheimer's Disease [72]. Tunnelling has also been seen in quantum biological tunnelling junctions [12]. Machine learning applications will involve designing algorithms that can efficiently harness the tunnelling process for healthcare applications. Quantum annealing is one area that allows efficient quantum machine learning through leveraging quantum tunnelling to enable effective algorithmic optimization [28, 30]. Screening and resolving genetic disease pathophysiology especially for the brain, which hosts nearly a trillion neurons, in an information network using quantum machine learning may benefit from methods using quantum annealing.

Quantum Entanglement and MachineLearning Applications for Medicine and Surgery

Entanglement is a quantum mechanical concept that Albert Einstein called "Spooky action at a distance" and vehemently rejected when it was first proposed [8, 11, 73]. This is where two particles become twinned and linked so that an action on one particle simultaneously affects the other particle's behavior and characteristics. Moreover, even when the particles are separated by great distances, Particle A's observables like momentum, spin, polarization, etc. are immediately reflected in Particle B's observables in perfect correlation when measured. What Einstein would not accept was that an intervention at one

place on Particle A can immediately affect the other Particle B as well, which he felt violated the principles of communicating at the speed of light.

Basic Mathematical Formalism of Entanglement Applied to Medicine [58, 74]

We will use pure states and Bra-Ket notation as we described above to illustrate a philosophical example. Let's consider two composite random quantum systems here, which we conjecture to represent a normal system and a diseased system with operators within the Hilbert Vector Space \mathbb{H}_{Norm} and $\mathbb{H}_{Disease}$.

These composite systems are represented by the Kronecker product \otimes of the two systems as described earlier in the chapter.

$$\mathbb{H}_{Norm} \otimes \mathbb{H}_{Disease}$$

$$\begin{aligned} \mathbb{H}_{Norm} &\text{ is in an initial state } |\psi_{Norm}\rangle \\ \mathbb{H}_{Disease} &\text{ is also in an initial state of } |\psi_{pre-Disease}\rangle \end{aligned}$$

An assumption here is that these are composite yet separable states, which can therefore be represented as:

$$|\psi_{Norm}\rangle \otimes |\psi_{pre-Disease}\rangle \text{ with their basis fixed such that we have } \{|i_{Norm}\rangle\} \text{ for } \mathbb{H}_{Norm} \text{ and } \{|j_{pre-Disease}\rangle\} \text{ for } \mathbb{H}_{Disease}$$

As well as the separability assumption for boundary conditions, which exists if \forall vectors like $Tissue_i^{Norm}$ and $Tissue_j^{pre-Disease}$, there exist at least one vector space wherein $Tissue_{i,j} \neq$

$Tissue_i^{Norm} Tissue_j^{pre-Disease}$ and also $Tissue_i^{Norm} \neq Tissue_j^{pre-Disease}$.

Logically it would make sense that within a universe of cells that make up the tissue encoded as a vector, there is a potential for at least one of these vector spaces to also exist where the above assumptions hold true.

As such this would be generalizable to:

$$|\psi_{Norm\&pre-Disease}\rangle = \sum_{ij} Tissue_{ij} (|i_{Norm}\rangle \otimes |j_{pre-Disease}\rangle).$$

Meaning that under such potential conditions, entanglement can therefore be defined for a given two fixed basis vectors as follows in a numerical example:

$$\{|0_{Norm}\rangle, |1_{Norm}\rangle\} \text{ for } \mathbb{H}_{Norm} \text{ and} \\ \{|0_{pre-Disease}\rangle, |1_{pre-Disease}\rangle\} \text{ for } \mathbb{H}_{Disease}$$

An entangled state could therefore be represented as follows:

$$\frac{1}{\sqrt{2}} (|0_{Norm}\rangle \otimes |1_{pre-Disease}\rangle - |0_{Norm}\rangle \otimes |1_{pre-Disease}\rangle)$$

This would suggest that in future it may be theoretically feasible to completely disentangle disease and normal states arising in the same setting to optimize treatment. Techniques may become available to study ways to potentially flip diseased states back to their normal states. However, while in-vitro this may be achievable using tools like Bose-Einstein condensates, etc., in-vivo there usually may be too much noise and interference to render these states undetectable, hidden or collapsed to a single observable. Quantum machine learning could facilitate identification and better delineation of these states, may enable finer control of the system's state and help manage decoherence, but would require innovations that allow noise-reduced multi-qubit computation at room temperatures.

Nonetheless, theoretically one can immediately see potential clinical usefulness for this interesting behavior of entanglement for various medical applications. Potential applications

include in the management of metastatic cancers, which are usually a poor prognostic marker for some of the most aggressive cancers known to man. Through the use of entanglement in medicine/surgery, one can attain the ability to diagnose and manage two location-separated yet identical disease processes simultaneously reducing treatment complication risks and side effects. In metastatic cancer management for instance, the ability to differentiate diseased-state from the non-diseased state and obtain the desired effect from chemoradiotherapy treatment could offer prolonged and improved clinical outcomes to extend life or reverse disease course. As an example, radiotherapy could be performed using entangled positron particles where its effect on one location of metastatic disease would be linked to another area with metastasis leading to uptake and simultaneous dose-calibrated treatment. In worst outcome tumors like glioblastoma multiforme, there is evidence of trans-hemispheric muscular tumor cell migration even prior to tumors being detectable [75]. Imaging of these tumors may not show their metastatic nature at other sites, which often look normal even to the expert eye through imaging. Since the exact mechanisms for how these tumors spread is still incompletely understood. In theory, quantum entangled machine learning systems could be used in medicine and surgical oncology as a way to not only predict and track this behavior to detect when these changes could occur/are occurring, but to facilitate effective treatment. Depending on the particle used for the computational modality one can even apply theranostic principles, where quantum algorithmically-guided nanoscale robotics can be implemented as illustrated in Fig. 10. This will allow complete access even to areas of the brain where conventional minimally invasive surgical methods are limited. Quantum machine learning could enable us to build better triangulation algorithms for simultaneous peri-operative highly targeted adjuvant and neoadjuvant chemoradiotherapy platforms. This then becomes useful depending on the location of the tumor and the eloquent tissue that may be affected, newer treatment modalities and more options will become available.

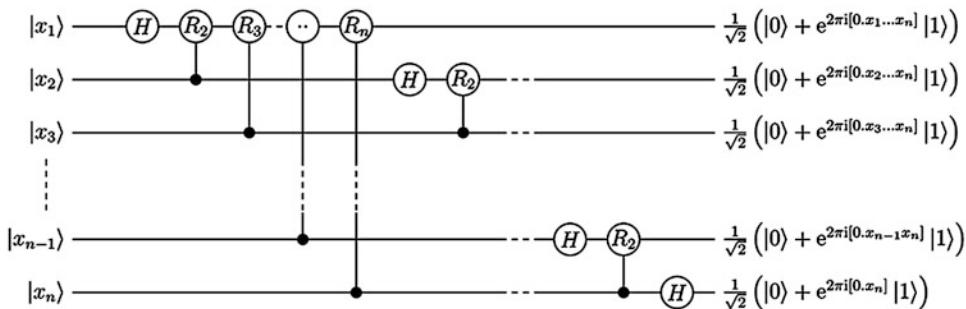


Fig. 9 Quantum circuit for Quantum-Fourier-Transform with n qubits (Image source from https://en.wikipedia.org/wiki/Quantum_Fourier_transform#/media/File:Q_fourier_nqubits.png)

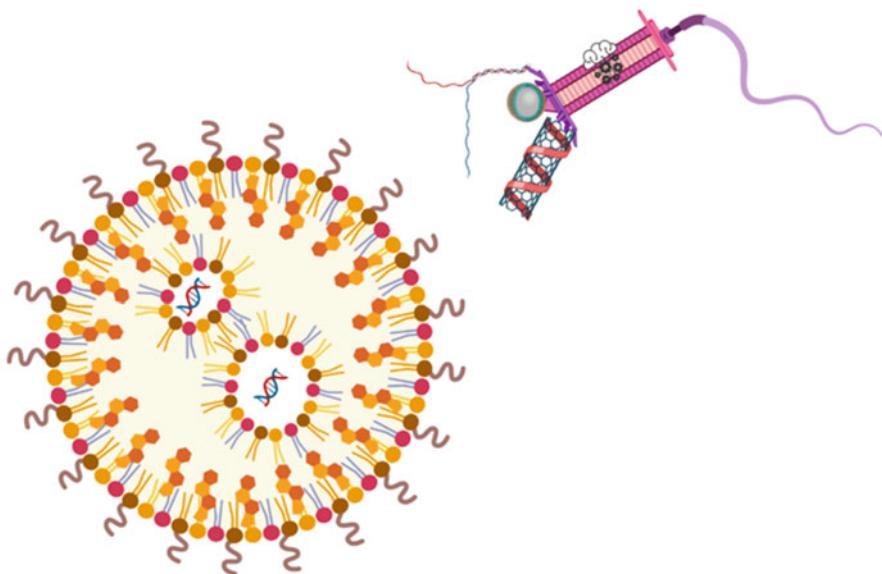


Fig. 10 Future Nanorobotic Surgery. Concept of a nano-scale robot capable of navigating towards and diagnosing the presence of a microbe or tumor cell, functionalized

with a nanotube, nanosphere, a gene molecule, and a qubit controllable computerized sensor. (Image created using BioRender)

Important Quantum Computing Algorithms and Quantum Machine Learning Algorithms

This section is not meant to be an exhaustive overview of all the algorithms, but is meant to guide the reader to some of the main quantum algorithms currently available for research and how they can be implemented or applied for future medical disease discovery and management. Quantum machine learning still remains a heavily theoretical field at present, but there are

simulators built to allow some practical implementations for research and experiments. Currently quantum algorithms have been designed to aid the optimization of tasks achieved by classical machine learning algorithms. Further current approaches also involve classical machine learning on quantum problems such as learning Hamiltonians. Open-source quantum computing has become assessable now and no longer confined to the laboratory [48–54, 58]. The circuitry of the algorithms is represented below for reference.

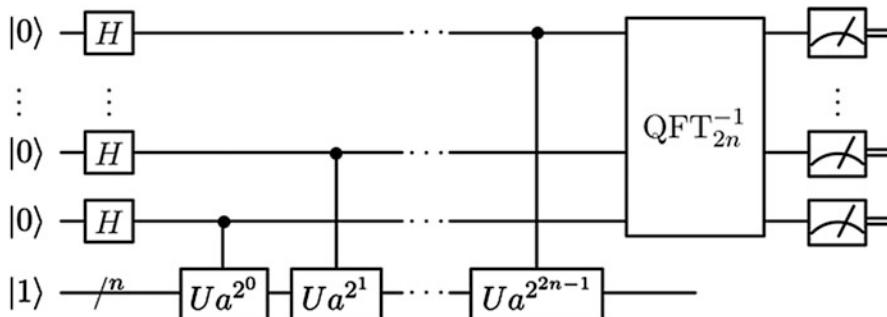


Fig. 11 Quantum subroutine in Shor’s algorithm for order finding, subsequently based on a Circuit for Shor’s algorithm using $2n+3$ qubits by Stephane Beauregard. Beauregard, Stéphane. “Circuit for Shor’s algorithm using $2n+3$

qubits.” Quantum Inf. Comput. 3 (2003): 175-185. (Image Source https://en.wikipedia.org/wiki/Shor%27s_algorithm#/media/File:Shor_s_algorithm.svg)

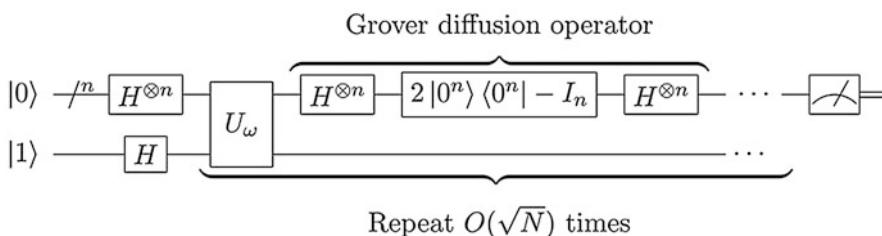


Fig. 12 A Quantum circuit representation of Grover’s algorithm (Image source by https://en.wikipedia.org/wiki/Grover%27s_algorithm#/media/File:Grover_s_algorithm_circuit.svg)

Quantum Fourier Transform

This algorithm invented by Dan Coppersmith is the classical discrete Fourier transform algorithm applied on a quantum state amplitude vector and is a linear transformation of qubits [76]. It forms the basis of several quantum algorithms useful for identifying the properties of qubits such as their phase in the phase estimation algorithm etc. This is useful for biomedical signal processing and also machine learning [73].

Shor’s Algorithm

This algorithm is a polynomial time integer factorization algorithm that was developed by Peter Shor to allow efficient prime factorization; a challenging feat for classical computers [77]. Its efficiency is linked to the efficiency of a Quantum Fourier Transform algorithm [73]. It was the first such

algorithm to demonstrate the vulnerability of RSA encryption to quantum computing. However, Shor’s algorithm fully depends on a functioning universal fault-tolerant quantum computation, which is yet to be made available in the NISQ era.

Grover’s Search Algorithm

An algorithm developed by Lov Grover that is able to probabilistically search through a database [78]. This quantum search algorithm is capable of finding a specific name among 100 million names in only 10,000 operations. It is characterized by Grover’s diffusion operator that is repeated over $O(\sqrt{N})$ time [11]. Quantum machine learning implementations of Grover’s search is being used to improve classical unsupervised learning algorithms like K-medians and K-nearest Neighbors and operate by amplitude augmentation of the wavefunction [79]. The classical equivalents

of these algorithms have been adopted in medicine for disease clustering and the quantum versions will improve disease pattern recognition and clustering as well as execution speed.

Deutsch-Jozsa Algorithm

This deterministic quantum algorithm proposed by David Deutsch and Richard Jozsa and one of the first of its kind to prove the exponential speed-up of a quantum computational algorithm [80]. It models the quantum computer as a black box to assess a function's ability to produce a balanced or constant deterministic output and poses the question of whether a *balanced* output (where the function returns 1 for half of the input domain and 0 for the other half) or a *constant* output of 0 for all outputs or 1 for all outputs can be returned. Pathologies linked to impaired homeostasis are common in the body, monitoring of homeostatic functions within the human body may benefit from the application and development of these algorithms (see Figure 13).

Bernstein-Vazirani Algorithm

Developed by David Bernstein and Umesh Vazirani, this is a restricted version of the Deutsch-Jozsa algorithm able to learn a string encoded function [24, 25]. This provides the most efficient function to identify a secret string of a particular length. Pathologies linked to genetic sequence abnormalities like missense mutations (e.g., sickle cell anemia) may benefit from the use of this algorithm (see Figure 14).

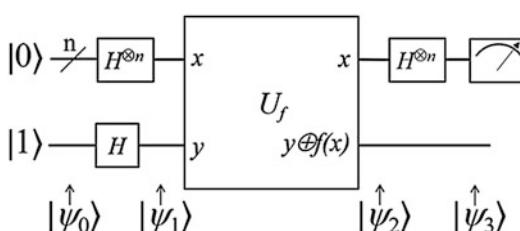


Fig. 13 Deutsch-Jozsa algorithm's quantum circuit (source from https://en.wikipedia.org/wiki/Deutsch%20%80%93Jozsa_algorithm#/media/File:Deutsch-Jozsa-algorithm-quantum-circuit.png)

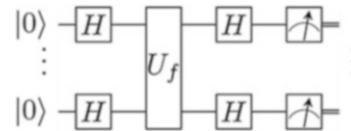


Fig. 14 Demonstrates a quantum circuit of the Bernstein-Vazirani algorithm (Image source https://en.wikipedia.org/wiki/Bernstein%20%80%93Vazirani_algorithm#/media/File:Bernstein-Vazirani_quantum_circuit.png)

Quantum Hidden Markov Chain Algorithm

These are the quantum equivalents to their classical hidden Markov chain algorithms for sequential and timeseries based modelling useful in various fields including robotics and natural language processing [81, 82]. These may be useful for modelling and predicting disease behavior related to stochastic probability distributions such as the randomness that exist in cancer cellular migration and tumor excision boundary conditions [75].

Quantum Natural Language Processing

The ability of computers to understand the higher-level information we give them is an important area of continuous study. University of Oxford's Robert Coecke, global collaborators and other institutions have been investigating applications of categorical quantum mechanics to computational linguistics and natural language processing [83]. Promising future extensions of this work to healthcare may include developing platforms to better model and aid the management of speech disorders for stroke, brain injuries and patients with mutism using quantum natural language processing algorithms and useful for improving AI-based radiology reporting systems.

Quantum Annealing and Quantum Neural Networks

Analogous to classical neural networks quantum neural networks are computational models based on quantum mechanics and aim to optimize sampling from higher dimensionality datasets and probability distributions [28, 30, 84]. It can also

enable an exponential augmentation in computational power compared to the classical computer.

Energy based models such as Boltzmann machines that obey the rules of entropy are usually very slow to train using classical computing. Multiple groups are exploring the use of quantum approaches to optimize the training of Boltzmann machines. Quantum annealing is able to generate samples from a Boltzmann distribution and is being generated as a potential to help train Boltzmann Machine [11, 73, 85]. NASA's D-Wave quantum system is one such example [53]. QNNs offer significant increases in computational power. They are deemed to be extensions of the Deutsch model of quantum computations. We have seen the tremendous potential for pattern recognition that neural networks are capable of and clinical practice will see its use-cases in improving computer vision systems image recognition and filter development for medical diagnostics.

Quantum Enhanced Reinforcement Learning

This distinct branch of classical machine learning adopts an agent-based approach and methodology to machine learning. Here the quantum agent interacts with the classical environment [73, 85]. Reinforcement learning using Quantum Boltzmann machines has been described recently in the literature [73, 85]. Here too there is a potential for developing improved quantum robotic platforms that can autonomously manage complex computations and navigate through the body without interference and will be useful for future nanorobotic surgery see Fig. 10 [86].

Quantum Phenomenon in Disease

Quantum biology was established in the 1920s and studies how the subatomic world of quantum mechanics plays a role in living cells. Current advances in research have identified phenomena like proton tunnelling, where the proton spontaneously disappears from one location within the cellular atom to another location, as a potential for

disease propagation [87, 88]. What allows such a phenomenon to occur is that the human DNA molecule is a helical structure with disulfide bridges and hydrogen atoms arranged at its periphery. These hydrogen atoms are protons and as a consequence could be observed in protonic exchanges across the helical structure. In fact, quantum proton tunnelling was identified as occurring at room temperature in DNA and has been hypothesized as a cause for cancer formation as this process can be propagated as mutations that arise during DNA replication [87]. Whether this contributes to an association with oncological disease states like Glioblastoma multiforme and others remains to be explored, but is an area of potential research enquiry.

Quantum Computing in Healthcare

In the healthcare industry, quantum computing could enable a range of disruptive use cases for providers and health plans by accelerating diagnoses through improved medical imaging analysis, personalized medicine, and optimized pricing. Quantum-enhanced machine learning algorithms are particularly relevant to the sector. Quantum computers are ideally suited for solving complex optimization tasks and performing fast searches of unsorted data. This is going to be relevant for many applications, from sorting and predicting the effects of climate data on health and disease or predicting healthcare financial data, to optimizing healthcare supply chain logistics, or global healthcare workforce management and deployment for disaster relief and robust integrations of healthcare ecosystems. Quantum encryption systems will offer us tremendous potential and safeguards for clinical data integrity. Quantum computing will also be useful for next generation drug discovery and reduction of pharmacokinetic and pharmacodynamic error, computational molecular optimizations and overall reducing false positive receptor interactions [12, 89–92]. Initiatives such as the planned National Quantum Computing Centre in Oxfordshire United Kingdom can foster a multidisciplinary environment for disease treatment discovery for healthcare, including optimised cancer

theranostics and complex genetic treatment discovery strategies.

Future Translational Considerations for E-Healthcare

The ability to simulate disease process and perform quantum Monte Carlo simulations will allow potential significant developments and discoveries in treating diseases or finding reversible paths for diseases in future. The simulation hypothesis and other controversial concept like the many universe construct would eventually be better studied and understood through quantum computing [93]. Quantum safety is another area where quantum computing and machine learning will benefit us greatly. With the quantum revolution, the safety of classical computers will be at risk. Cybersecurity and newer forms of cyberwarfare against e-healthcare are also areas where one would need quantum computers to safeguard against leading to a trend towards absolute security and via Quantum Diffie-Hellman and Quantum Group Key Agreements and other systems [94, 95]. Other areas such as the already demonstrated quantum teleportation will also allow us to establish methods to achieving healthcare information transfer across far greater interplanetary distances and perhaps teleportation of man will not be a far-fetched science fictional concept that will present its own medical challenges to overcome.

Currently, the issues affecting effective design and operations of quantum computational systems include quantum decoherence of the definite phase relationship between all the superposed states [96]. This is attributed to systematic noise and work to reduce quantum noise has made significant progress in the recent years. However, methods for decoherence needs to be overcome and fully understood for the human body as a quantum system in order to fully unlock applications of ensembles quantum computation and building devices like effective quantum efficient brain machine interfaces. There's a belief that this will be achieved in the distant future. Through better tools for quantum computation, there will also be a better understanding and delineation of

various forms of fuzzy logic from quantum logic and provide medically relevant applications. In the near term, we see small qubit quantum computers aiding boosting machine learning algorithms. However, it remains unclear about the timeline for most of these future developments outlined above and would depend on funding avenues for translational collaboration between government, industry and academic institutions to benefit healthcare.

Ethico-Legal Implications and Considerations

Insurance systems for healthcare economies look at the potential risk of individuals developing conditions and uses that to forecast whether monthly payment contributions are appropriate for that individual. From one perspective individuals might be forced to pay more based on the predictive use of quantum and machine learning methodologies for probabilistic disease course prediction. This opens up new regulatory challenges. Consider a scenario where an individual with a high probability of a genetic disorder is identified. Not only will it have repercussions for information and data protection, stigma control and psychological burden control, but also perhaps may cause undue distress to family particularly if there is only a single potential avenue of treatment for them [97].

One would counter-argue that this same individual will be helped if we could predict the probabilities of various disease states and the pathophysiological patterns that would lead to an unfavorable outcome. This would guide lifestyle choices and provide treatment avenues that would lead to better outcomes for them and hence the model of healthcare insurance will need to change to support or factor in this process.

There is an argument for both, and multidisciplinary consortia of computer scientists, physicists, chemists, mathematicians, economists, medics and policy-makers will need to be established in order to enable consensus decisions around medical use-cases of quantum computation and machine learning.

Summary Remarks

It is conceivable that quantum computers have significant medical potential. From the ability to improve diagnostics through the analysis of medical images, including processing steps, such as pattern recognition and image matching, to improving the medical image acquisition speed that can be accelerated using quantum machine learning algorithms. These use-cases as well as those described above will significantly help advance healthcare's quadruple aims. The combination with AI will not only make patient care more efficient leading to improved population health, improved quality of life and earlier diagnoses, such as diagnosing and perhaps reversing cancer at an early stage, but will truly provide the quantum leap for mankind. However, significant challenges still remain, but the medical community has a lot of positive areas to inspire research enquiry.

Nearly 200 years after Charles Babbage proposed his innovative idea of the “difference engine” and it was subsequently prototyped, we may finally be making the leap towards realizing his dream through quantum computing.

References

1. Braibant S, Giacomelli G, Spurio M. Particles and fundamental interactions: an introduction to particle physics. 2nd ed. Springer; 2012.
2. Phillips R. In retrospect: the Feynman lectures on physics. *Nature*. 2013;504(7478):30–1.
3. Albrecht A, et al. Self-assembling hybrid diamond–biological quantum devices. *New J Phys*. 2014;169: 093002.
4. Calafell A, et al. Quantum computing with graphene plasmons. *npj Quantum Inf*. 2019;5:37. <https://doi.org/10.1038/s41534-019-0150-2>.
5. Chawla P, Chandrashekar CM. Quantum walks in polycyclic aromatic hydrocarbons. *arXiv:201214463v1 [quant-ph]*. 2020.
6. Shim J, et al. Robust dynamical decoupling for arbitrary quantum states of a single NV center in diamond. *EPL*. 2012;99:40004. <https://doi.org/10.1209/0295-5075/99/40004>.
7. McKie R. Carbon Capture Vodka, toothpaste, yoga mats . . . the new technology making items out of thin air. *Guardian Newspaper*, 2021.
8. Simon C. Natural entanglement in Bose-Einstein condensates. *Phys Rev A*. 2002; 665.
9. Watson JD, Crick F. A structure for deoxyribose nucleic acid. *Nature*. 1953;171:737–8.
10. Deaton R. DNA and quantum computers. *GECCO'01: Proceedings of the 3rd annual conference on genetic and evolutionary computation July 2001*. 2001. p. 989–96.
11. Mohamed K. Neuromorphic computing and beyond: parallel, approximation, near memory, and quantum. Springer Nature; 2020.
12. Xin H, Sim WJ, Namgung B, Choi Y, Li B, Lee LP. Quantum biological tunnel junction for electron transfer imaging in live cells. *Nat Commun*. 2019;10(1):3245.
13. Zhang Y, Wang F, Chao J, et al. DNA origami cryptography for secure communication. *Nat Commun*. 2019;10:5469. <https://doi.org/10.1038/s41467-019-13517-3>.
14. Panda D, Molla KA, Baig MJ, Swain A, Behera D, Dash M. DNA as a digital information storage device: hope or hype? *3 Biotech*. 2018;8(5): 239. <https://doi.org/10.1007/s13205-018-1246-7>. PMC 5935598
15. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J*. 1948;27(3):379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>; hdl:10338.dmlcz/101429
16. Chiribella G, Kristjánsson H. Quantum Shannon theory with superpositions of trajectories. *Proc R Soc A*. 2018. <https://doi.org/10.1098/rspa.2018.0903>.
17. Collier B, MacLachlan J. Charles Babbage: and the engines of perfection. Oxford University Press; 2000. p. 29–30. ISBN 978-0-19-514287-7.
18. Hegade N, et al. Experimental demonstration of quantum tunneling in IBM quantum computer. *arXiv:171207326v4:1-42*. 2019.
19. Conover E. The new light-based quantum computer Jiuzhang has achieved quantum supremacy. *Science News*. 2020. Retrieved December 07, 2020.
20. Arute F, Arya K, Babbush R, et al. Quantum supremacy using a programmable superconducting processor. *Nature*. 2019;574:505–10. <https://doi.org/10.1038/s41586-019-1666-5>.
21. Rosser J. An informal exposition of proofs of Gödel's theorem and Church's theorem. *J Symb Log*. 1939;4 (2):53–60. <https://doi.org/10.2307/2269059>. JSTOR 2269059.
22. Deutsch D. Quantum theory, the Church-Turing principle and the universal quantum computer. *Proc R Soc A*. 1985;400(1818):97–117.
23. Benioff P. Quantum mechanical hamiltonian models of turing machines. *J Stat Phys*. 1982;29(3):515–46.
24. Falalek S, Herold CD, McMahon BJ, Maller KM, Brown KR, Amini JM. Transport implementation of the Bernstein–Vazirani algorithm with ion qubits. *New J Phys*. 2016; 18. <https://doi.org/10.1088/1367-2630/aab341>.
25. Bernstein E, Vazirani U. Quantum complexity theory. *SIAM J Comput*. 1997;26(5):1411–73. <https://doi.org/10.1137/S0097539796300921>.
26. Ball P. Physicists in China challenge Google's ‘quantum advantage’. *Nature*. 2020;588(7838):380.

27. Steane A. The ion trap quantum information processor. *Appl Phys B Lasers Opt.* 1997;64:623–42.
28. Kadovski T, Nishimori H. Quantum annealing in the transverse Ising model. *Phys Rev E.* 1998;58:5355.
29. Farhi E, et al. A quantum adiabatic evolution algorithm applied to random instances of an NP-complete problem. *Science.* 2001;292:472–5.
30. Li RY, Di Felice R, Rohs R, et al. Quantum annealing versus classical machine learning applied to a simplified computational biology problem. *npj Quantum Inf.* 2018;4:14. <https://doi.org/10.1038/s41534-018-0060-8>.
31. Ivády V, Davidsson J, Delegan N, et al. Stabilization of point-defect spin qubits by quantum wells. *Nat Commun.* 2019;10:5607. <https://doi.org/10.1038/s41467-019-13495-6>.
32. Saffman M. Quantum computing with neutral atoms. *Natl Sci Rev.* 2019;6(1):24–5.
33. Lahtinen V, Pachos J. A short introduction to topological quantum computation. *arXiv: Mesoscale and Nanoscale Physics.* 2017.
34. Tan S-H, Rohde PP. The resurgence of the linear optics quantum interferometer – recent advances & applications. *Rev Phys.* 2019;4:100030.
35. Watson T, Philips S, Kawakami E, et al. A programmable two-qubit quantum processor in silicon. *Nature.* 2018;555:633–7. <https://doi.org/10.1038/nature25766>.
36. Ramamoorthy A. Switching characteristics of coupled quantum wires with tunable coupling strength. *Appl Phys Lett.* 2006;89:013118. <https://doi.org/10.1063/1.2219085>.
37. Qiu X, Zou J, Qi X, et al. Precise programmable quantum simulations with optical lattices. *npj Quantum Inf.* 2020;6:87. <https://doi.org/10.1038/s41534-020-00315-9>.
38. Ansaldi F, Chatterjee A, Bohuslavskyi H, Bertrand B, Hutin L, Vinet M, et al. Single-electron operations in a foundry-fabricated array of quantum dots. *Nat Commun.* 2020. <https://doi.org/10.1038/s41467-020-20280-3>.
39. Cory D, Fahmy A, Havel T. Ensemble quantum computing by NMR spectroscopy. *Proc Natl Acad Sci.* 1997;94(5):1634–9. <https://doi.org/10.1073/pnas.94.5.1634>.
40. Kane BE. A silicon-based nuclear spin quantum computer. *Nature.* 1998;393:133.
41. Badrutdinov A, et al. Nonlinear transport of the inhomogeneous Wigner solid in a channel geometry. *Phys Rev B.* 2016. <https://doi.org/10.1103/PhysRevB.94.195311>.
42. Ohlsson N, Krishna Mohan R, Kröll S. Quantum computer hardware based on rare-earth-ion-doped inorganic crystals. *Opt Commun.* 2002;201(1):71–7.
43. Blais A, Girvin SM, Oliver WD. Quantum information processing and quantum optics with circuit quantum electrodynamics. *Nat Phys.* 2020;16:247–56. <https://doi.org/10.1038/s41567-020-0806-z>.
44. Náfrádi B, Chouair M, Dinse KP, et al. Room temperature manipulation of long lifetime spins in metallic-like carbon nanospheres. *Nat Commun.* 2016;7:12232. <https://doi.org/10.1038/ncomms12232>.
45. Ju C, Suter D, Du J. An endohedral fullerene-based nuclear spin quantum computer. *Phys Lett A.* 2011;375(12):1441–4.
46. Bradley CE, Randall J, Abobeih MH, Berrevoets RC, Degen MJ, Bakker MA, et al. A ten-qubit solid-state spin register with quantum memory up to one minute. *Phys Rev X.* 2019;9(3):031045.
47. Andrianov SN, Moiseev SA. Magnon qubit and quantum computing on magnon Bose-Einstein condensates. *Phys Rev A.* 2014;90(4):042303.
48. Microsoft. Introduction to Azure Quantum Online: Microsoft; 2020. Available from: <https://docs.microsoft.com/en-us/azure/quantum/overview-azure-quantum>
49. AtomComputing. Atom Computing: Atom Computing. Available from: <https://www.atom-computing.com/>
50. XanaduQuantum. Xanadu Quantum Cloud. 2021. Available from: <https://www.xanadu.ai/>
51. IBM. IBM Quantum Experience. 2021. Available from: <https://quantum-computing.ibm.com/>
52. ColdQuanta. Cold Quanta Quantum Computing. 2021. Available from: <https://coldquanta.com/>
53. DWave. D-Wave Quantum Computing. Dwave; 2021. Available from: <https://www.dwavesys.com/quantum-computing>
54. StrangeWorks. Strange Works Quantum Computing. 2021. Available from: <https://strangeworks.com/>
55. Hou S-y, et al. SpinQ Gemini: a desktop quantum computer for education and research. *arXiv:210110017v2 [quant-ph].* 2021.
56. Drexler KE. Reframing superintelligence: comprehensive AI services as general intelligence. Technical Report #2019-1. Future of Humanity Institute, University of Oxford; 2019. https://www.fhi.ox.ac.uk/wp-content/uploads/Reframing_Superintelligence_FHI-TR-2019-1.1-1.pdf
57. Sandberg A, Bostrom N. Whole brain emulation: a roadmap. Technical Report #2008-3. Future of Humanity Institute, Oxford University; 2008. www.fhi.ox.ac.uk/reports/2008-3.pdf
58. Scherer W. Mathematics of quantum computing. Springer Nature Switzerland AG; 2019.
59. Ambainis A. What can we do with a quantum computer? How quantum information could lead to a better understanding of the principles of all quantum systems. Institute of Advanced Study; 2014. Available from: <https://www.ias.edu/ideas/2014/ambainis-quantum-computing>
60. Sen D. The uncertainty relations in quantum mechanics. *Curr Sci.* 2014;107(2):203–18.
61. Nimtz G. Tunneling violates special relativity. *arXiv:10033944.* 2010.
62. Ciaglia FM, Ibort A, Marmo G. Schwinger's picture of quantum mechanics I: groupoids. *Int J Geom Meth Mod Phys.* 2019;1608:1950119.

63. Styer D, et al. Nine formulations of quantum mechanics. *Am J Phys.* 2002;70:288–97. <https://doi.org/10.1119/1.1445404>.
64. Solenov D, et al. The potential of quantum computing and machine learning to advance clinical research and change the practice of medicine. *Mo Med.* 2018;115(5):463–7.
65. Musk E, Neuralink. An integrated brain-machine interface platform with thousands of channels. *J Med Internet Res.* 2019;21(10):e16194. <https://doi.org/10.2196/16194>. PMID: 31642810 PMCID: 6914248.
66. Spielmann C, Szipoes R, Stingl A, Krausz F. Tunneling of optical pulses through photonic band gaps. *Phys Rev Lett.* 1994;73:2308–11.
67. Steinberg A, Kwiat PG, Chiao RY. Measurement of the single-photon tunneling time. *Phys Rev Lett.* 1993;71:708–11.
68. Sekatskii S, Letokhov V. Electron tunneling time measurement by field-emission microscopy. *Phys Rev B.* 2001;64:233311, 1–4.
69. Eckle P, Pfeiffer A, Cirelli C, Staudte A, Dörner R, Muller H, et al. Attosecond ionization and tunneling delay time measurements in helium. *Science.* 2008;322:1525–9.
70. Yang S, Page J, Liu Z, Cowan M, Chan C, Sheng P. Ultrasound tunneling through 3D phononic crystals. *Phys Rev Lett.* 2002;88:104301, 1–4.
71. Robertson W, Ash J, McGaugh J. Breaking the sound barrier: tunneling of acoustic waves through the forbidden transmission region of a one-dimensional acoustic band gap array. *Am J Phys.* 2002;70:689–93.
72. Lee J, Kang DY, Kim SU, Yea CH, Oh BK, Choi JW. Electrical detection of beta-amyloid (1–40) using scanning tunneling microscopy. *Ultramicroscopy.* 2009;109(8):923–8. <https://doi.org/10.1016/j.ultramic.2009.03.009>. Epub 2009 Mar 19.
73. Sarma S, Deng DL, Duan L-M. Machine learning meets quantum physics. *Phys Today.* 2019;48–54.
74. Brilliant.org. Quantum Entanglement. 2021. Retrieved 16:54, April 29, 2021, from <https://brilliant.org/wiki/quantum-entanglement/>
75. Holland E. Glioblastoma multiforme: the terminator. *Proc Natl Acad Sci U S A.* 2000;97(12):6242–4. <https://doi.org/10.1073/pnas.97.12.6242>.
76. Coppersmith D. An approximate Fourier transform useful in quantum factoring. Technical Report RC19642, IBM. 1994.
77. Shor PW. Algorithms for quantum computation: discrete logarithms and factoring. In: Proceedings 35th annual symposium on foundations of computer science. IEEE Comput Soc Press; 1994. p. 124–34. <https://doi.org/10.1109/sfcs1994365700>. ISBN 0818665807.
78. Grover LK. A fast quantum mechanical algorithm for database search. In: Proceedings of the twenty-eighth annual ACM symposium on Theory of Computing STOC '96 Philadelphia, Pennsylvania, USA. Association for Computing Machinery; 1996. p. 212–9. <https://doi.org/10.1145/237814.237866>. ISBN 978-0-89791-785-8.
79. Aïmeur E, Brassard G, Gambs S. Quantum clustering algorithms. In: Proceedings of the 24th international conference on Machine learning; Corvalis, Oregon, USA. Association for Computing Machinery; 2007. p. 1–8.
80. Deutsch D, Jozsa R. Rapid solutions of problems by quantum computation. *Proc R Soc Lond A.* 1992;439(1907):553–8.
81. Clark LA, et al. Hidden quantum Markov models and open quantum systems with instantaneous feedback. In: Emergence, complexity and computation. 2015. p. 143–51. 2014.
82. Cholewa M, et al. Quantum hidden Markov models based on transition operation matrices. *Quantum Inf Process.* 2017;16:1–19.
83. Lorenz R, et al. QNLP in practice: running compositional models of meaning on a quantum computer. *ArXiv abs/210212846.* 2021.
84. Gupta S, Zia R. Quantum neural networks. *J Comput Syst Sci.* 2001;63(3):355–83.
85. Crawford D, Levit A, Ghadermarzy N, Oberoi J, Ronagh P. Reinforcement learning using quantum Boltzmann machines. *arXiv:161205695 [quant-ph].* 2018.
86. Li J, Esteban-Fernandez de Avila B, Gao W, Zhang L, Wang J. Micro/nanorobots for biomedicine: delivery, surgery, sensing and detoxification. *Sci Robot.* 2017; 2(4). <https://doi.org/10.1126/scirobotics.aam6431>.
87. Srivastava R. The role of proton transfer on mutations. *Front Chem.* 2019;7(536).
88. Pusuluk O, Farrow T, Deliduman C, Burnett K, Vedral V. Proton tunnelling in hydrogen bonds and its implications in an induced-fit model of enzyme catalysis. *Proc R Soc A: Math Phys Eng Sci.* 2018;474(2218): 20180037.
89. Kotev M, Sarrat L, Gonzalez CD. User-friendly quantum mechanics: applications for drug discovery. *Methods Mol Biol.* 2020;2114:231–55. https://doi.org/10.1007/978-1-0716-0282-9_15.
90. Lodola A, De Vivo M. The increasing role of QM/MM in drug discovery. *Adv Protein Chem Struct Biol.* 2012;87:337–62. <https://doi.org/10.1016/B978-0-12-398312-1.00011-1>.
91. Bryce R. What next for quantum mechanics in structure-based drug discovery? *Methods Mol Biol.* 2020;2114:339–53. https://doi.org/10.1007/978-1-0716-0282-9_20.
92. Thomford N, Senthebene D, Rowe A, Munro D, Seele P, Maroyi A, et al. Natural products for drug discovery in the 21st century: innovations for novel drug discovery. *Int J Mol Sci.* 2018;19(6):1578. <https://doi.org/10.3390/ijms19061578>. PMID: 29799486; PMCID: PMC6032166.
93. Ashrafi H. How many simulations do we exist in? A practical mathematical solution to the simulation argument. *arXiv: Pop Phys.* 2020.

94. Naresh V, Nasralla MM, Reddi S, García-Magariño I. Quantum Diffie-Hellman extended to dynamic quantum group key agreement for e-healthcare multi-agent systems in smart cities. *Sensors (Basel)*. 2020;20(14):3940. <https://doi.org/10.3390/s20143940>. PMID: 32679823; PMCID: PMC7412309.
95. Abd-El-Atty B, Iliyasu AM, Alaskar H, Abd El-Latif AA. A robust quasi-quantum walks-based steganography protocol for secure transmission of images on cloud-based E-healthcare platforms. *Sensors (Basel)*. 2020;20(11):3108.
96. Schreier J, et al. Suppressing charge noise decoherence in superconducting charge qubits. *Phys Rev B*. 2008;77:180502. <https://doi.org/10.1103/PhysRevB77180502>, arXiv:07123581.
97. O'Neil C. Weapons of math destruction: how big data increases inequality and threatens democracy. Crown Publishing Group; 2016.
98. Caruso F, Crespi A, Ciriolo A, et al. Fast escape of a quantum walker from an integrated photonic maze. *Nat Commun.* 2016;7:11682. <https://doi.org/10.1038/ncomms11682>.

Part III



Emergence of Deep Machine Learning in Medicine

31

Richard Dybowski

Contents

Introduction	449
Deep Neural Networks	450
The Universal Approximation Theorem and Its Limitation	450
Internal Hierarchical Feature Extraction	450
Internal Transformation of Dataset Topologies	452
Medical Examples	455
Medical Imaging	455
Genomics and Epigenomics	455
Natural Language Processing	455
Conclusion	456
References	456

Abstract

Deep learning has been shown to be of benefit in a number of areas of medicine. This chapter gives a recent topological explanation of the efficacy of deep learning in terms of internal transformations of dataset topologies by deep neural networks. This is followed by medical examples including imaging and genomics.

Introduction

Of the various machine learning techniques that have been applied to medical problems, the use of neural networks has arguably been the most prominent, and this has been the case since the early 1990s [3] (This chapter assumes that the reader is familiar with the basic concepts associated with neural networks as detailed by Bishop [4, 5].).

Before the 1990s, the main AI approach to clinical decision-making was to attempt to use knowledge-based systems [20]. These systems supposedly encapsulated medical expert knowledge, but the development of these so-called expert systems was beset with difficulties [44];

R. Dybowski (✉)
St John's College, Cambridge, UK
e-mail: rd460@cam.ac.uk

consequently, attention turned to data-based approaches, and to machine learning in particular.

Rosenblatt [38] developed the Perceptron – a neural network without a hidden layer – but it was limited to classifying feature vectors that are linearly separable [32], a limitation that brought about the first “AI Winter.” It was not until the publication of the back-propagation algorithm by Rumelhart et al. [39] for the training of neural networks with a hidden layer that an explosion of interest in neural computing took place (However, the back-propagation algorithm was first discovered by Werbos [45].).

Despite the interest in neural networks that occurred during the 1980s and 1990s, frustrations were also present partly because of limitations with the methodology used at the time; for example, the use of sigmoidal activation functions such as the hyperbolic tanh. It were improvements such as the introduction of the ReLU activation function [33] to overcome the vanishing gradient problem [16], and the use of the “dropout” method to suppress overfitting [42], that helped to bring neural networks to the standard they are today and establish the renaissance in neural computation ($\text{ReLU}(x) = \max(0, x)$. ; Coupled with the availability of faster computers and more data.).

Deep Neural Networks

The terms “shallow” and “deep,” when applied to neural networks, simply refer to the depth of a network: A shallow neural network has no more than one hidden layer whereas a *deep neural network* has several hidden layers.

Consider the data points displayed in Fig. 1a. These are *linearly separable*, which means that the classes can be completely separated by a single (hyper)plane between them. Consequently, a linear discriminant function [15] such as a logistic regression model (Fig. 2) can be considered as a classifier with respect to the data. In contrast, the distribution of the data in Fig. 1b is not linearly separable; however, note that a nested regression model (Fig. 3) is able to produce nonlinear decision boundaries.

A single-layer perceptron (Fig. 4) with a logistic activation function is equivalent to a logistic regression model; and a feed-forward neural network (FNN) with a single hidden layer can be viewed as a nested logistic model, but the analysis can be taken further as follows.

The Universal Approximation Theorem and Its Limitation

Cybenko [6] and Hornik [17] proved that an FNN with a sufficiently large single hidden layer can, in theory, approximate any continuous function (the *Universal Approximation Theorem*). Although this is true, the hidden layer may have to be extremely large in order to achieve this for a given task, and the larger the hidden layer, the greater the number of weights (parameters) required for that layer. Moreover, determining the number of nodes needed in that hidden layer for a given task is difficult. On the other hand, as argued by Mhaskar et al. [31], fewer parameters are required when more than one hidden layer is used.

Internal Hierarchical Feature Extraction

In the context of statistics and machine learning, the term *feature extraction* refers to the derivation of new values (features) from measured data to facilitate the analysis of the data and/or perform predictions. And, when used in the context of deep neural networks, the phrase “hierarchical feature extraction” is primarily applied to convolutional neural networks.

Convolutional neural networks (CNNs) are spatially invariant deep networks that use convolution in place of the matrix algebra of feed-forward networks in at least one of their layers [12]. The architecture of a CNN, originally proposed by LeCun [25], was inspired by the Hubel and Wiesel [18] model of the visual primary cortex.

The *convolution* of a layer of values within a CNN uses local kernels to detect local features.

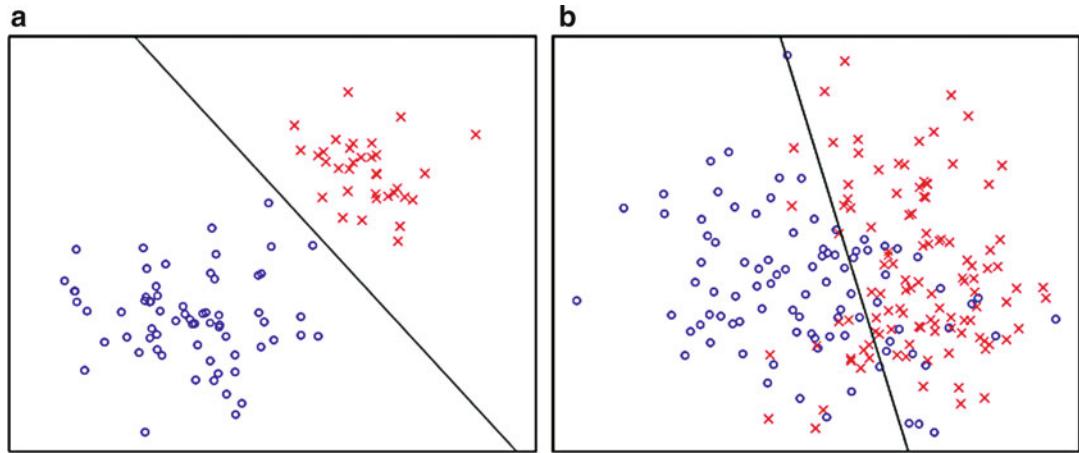


Fig. 1 Class-labeled data points: (a) linearly separable and (b) not linearly separable. The line in (a) provides a decision boundary for classification

$$\mathcal{E}[Y|x_1, x_2] = \frac{1}{1 + \exp [-(w_0 + w_1x_1 + w_2x_2)]}$$

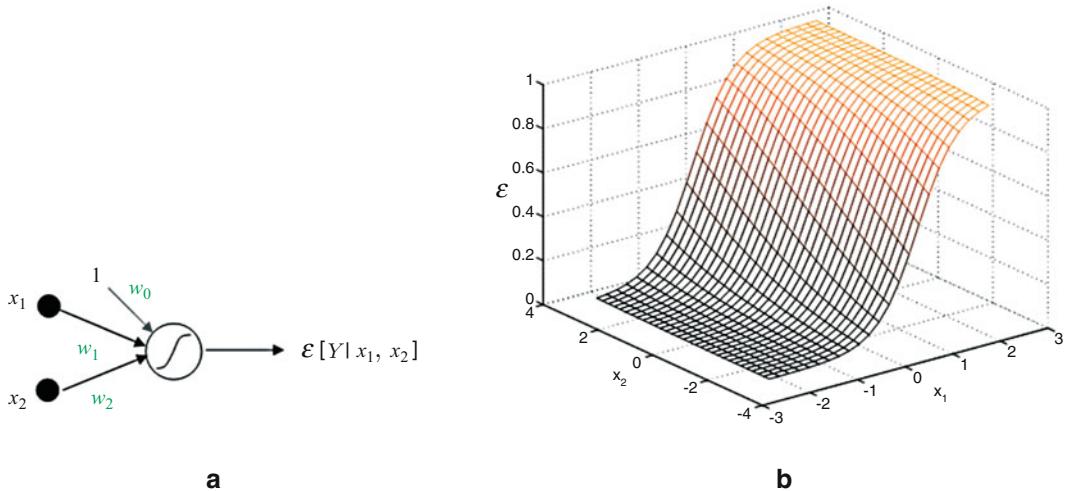


Fig. 2 Conditional expectation $\mathcal{E}[Y|x_1, x_2]$ is equal to the logistic function $\text{Logistic}(x_1, x_2; w_1, w_2, w_3)$. Classification can be based on whether $\mathcal{E}[Y|x_1, x_2]$ is greater than some

threshold value τ (e.g., $\tau = 0.5$), and the resulting decision boundary (not shown) will be linear

This results in a type of feature map at the next layer. Another technique called *max pooling* reduces the dimensionality of a feature map but retains the most important information to be passed on to the following layer. A common architecture for CNNs is to have alternate

convolution and pooling layers, with a shallow feed-forward neural network for the final layers to provide a probability output.

The use of convolution and max pooling produces a successive abstraction of information held within a series of layers. In their [26] paper, in

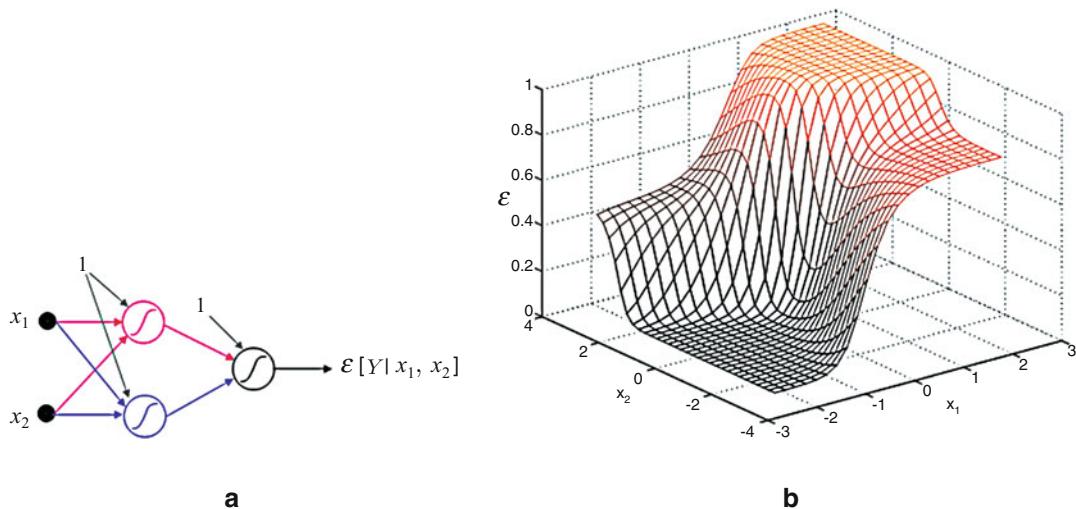


Fig. 3 A nested logistic regression model, which provides a nonlinear decision boundary for classification

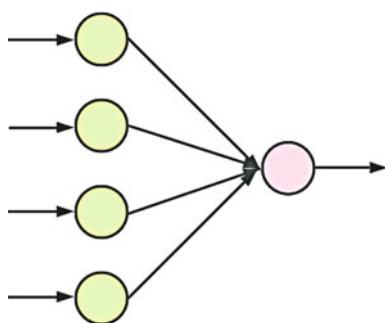


Fig. 4 FNN with no hidden layers (a “perceptron”): $Net(\mathbf{x}) = f_{out}(\mathbf{x}) = Logistic(\mathbf{x}; \mathbf{w})$, where \mathbf{x} is the vector of input values and \mathbf{w} are the weights

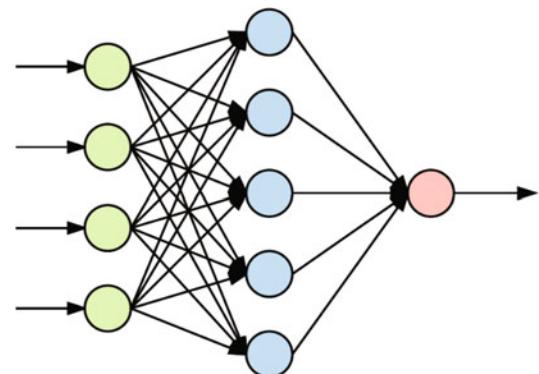


Fig. 5 FNN with one hidden layer (blue): $Net(\mathbf{x}) = f_{out} \circ f_1(\mathbf{x})$

which they focused on CNNs, LeCun et al. wrote “Deep neural networks exploit the property that many natural signals are compositional hierarchies, in which higher-level features are obtained by composing lower-level ones.” Each layer in the network can be thought of as performing an abstraction of the information held within the preceding layer so that a sequence of layers provides a hierarchy of increasing abstraction. This can obviate the need for the manual selection of input features once required for most neural networks [43] (Fig. 5).

Internal Transformation of Dataset Topologies

The deep FNN, $Net(\mathbf{x})$, shown in Fig. 6 can be deconstructed into two FNNs. One FNN, $Net_1(\mathbf{x})$, has all the hidden layers but not the output node (i.e., $Net_1(\mathbf{x}) = f_3 \circ f_2 \circ f_1(\mathbf{x})$), the other FNN is just the logistic output node function f_{out} ; consequently, $Net(\mathbf{x}) = f_{out} \circ Net_1(\mathbf{x})$.

A mathematical manifold is a topological space that locally resembles Euclidean space near each point, and there is evidence that real-

Fig. 6 FNN with three hidden layers: $Net(\mathbf{x}) = f_{out} \circ f_3 \circ f_2 \circ f_1(\mathbf{x})$. Note that there can be more than four input nodes as well as more than one output node and many hidden layers in an FNN

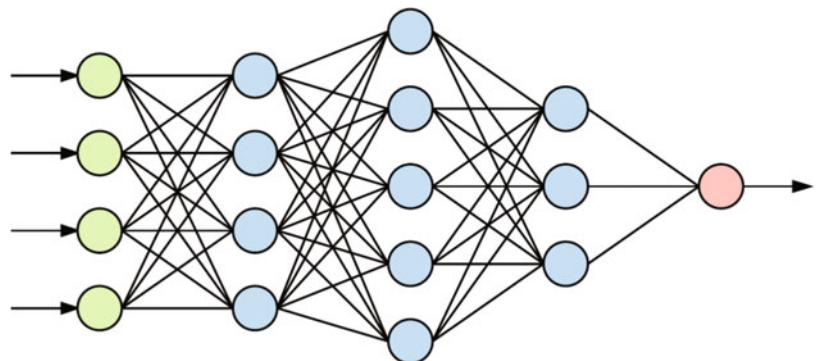
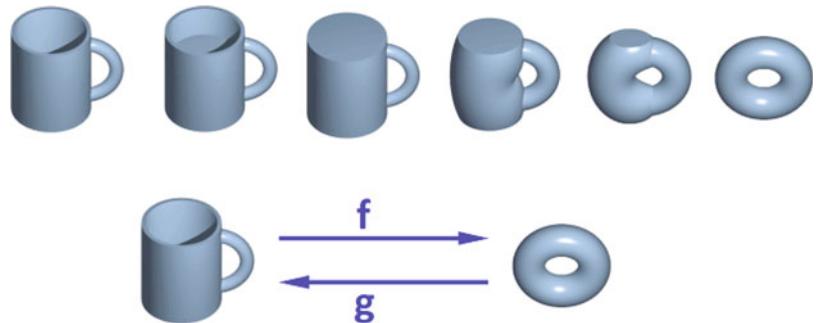


Fig. 7 Homeomorphism f of a mug to a doughnut. g is the inverse



world data points exist on manifolds [10] – the *manifold hypothesis*. Topology is a branch of mathematics concerned with the properties of geometric objects that are preserved under continuous deformations, such as stretching, twisting, crumpling, and bending, but not tearing or gluing [1], and Olah [35] looked at the efficacy of neural networks topologically. He proposed that a deep FNN $Net(\mathbf{x})$ learns the topological transformation required to create a homeomorphism of the original data manifold M such that the mappings of the data points on M onto the resulting manifold via $Net_1(\mathbf{x})$ are linearly separable according to f_{out} (A *homeomorphism* is a continuous function between topological spaces that has a continuous inverse function. An example is shown in Figure 7.). The inverse Net^{-1} creates a nonlinear decision boundary for classification (Fig. 9).

As an example, consider the two sets of data (curves) shown in Fig. 8a, which are on a simple 2D manifold. These are clearly not linearly separable; however, a homeomorphic distortion of the

manifold provides linear separability (Figs. 8b and 9).

Lei et al. [27] posited that the fundamental principle attributing to the success of deep learning is the manifold structure in data, namely that natural high-dimensional data concentrates close to a low-dimensional manifold, and deep learning learns the manifold and the probability distribution on it.

Detailed analysis of deep learning using topology has recently been conducted by Naitzat et al. [34] and Hajij and Istvan [14]. Naitzat et al. used Betti numbers [2] for their analysis and showed that deep networks work by reducing the Betti numbers of manifolds M . They made the following observations:

- **Effects of depth on topology change:** Reducing the depth of a constant-width network beyond a certain threshold makes it increasingly difficult to train the network to high accuracy. Moreover, as the depth is reduced,

Fig. 8 (a) Two curves on a manifold, which are not linearly separable. (b) Curves linearly separable after the manifold is transformed by an FNN [35]

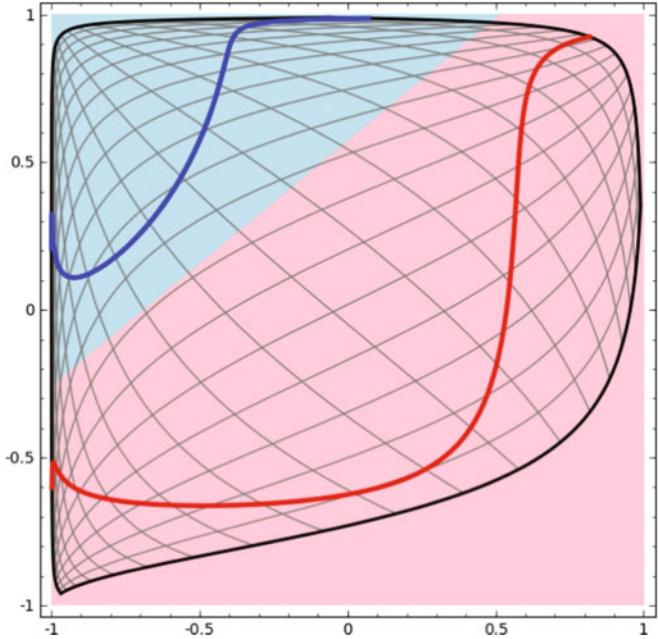
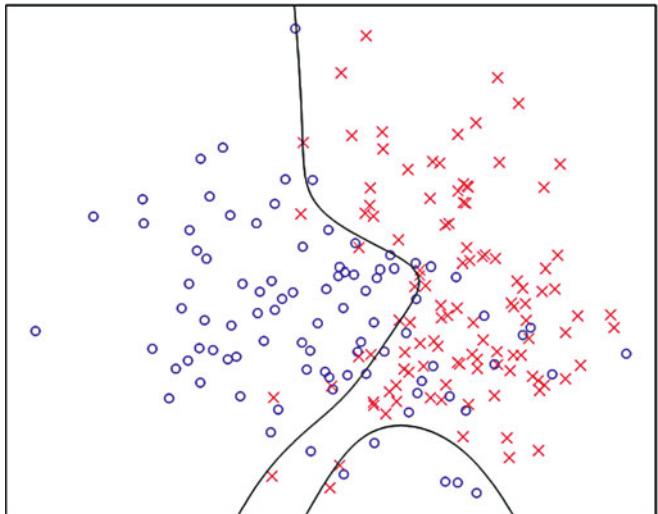


Fig. 9 Classification decision boundary for Fig. 1b obtained from an FNN



the burden of changing topology does not spread evenly across all hidden layers but becomes concentrated in the final layers. The initial layers do not appear to play a big role in changing topology, and reducing depth simply makes the final layers “work harder” to produce larger reductions in Betti numbers.

- **Nonhomeomorphic activations induce rapid topology changes:** The tanh activation

function is less effective at reducing Betti numbers, whereas the nonhomeomorphic activation ReLU exhibits the most rapid reductions in Betti numbers. The effectiveness of ReLU over sigmoidal activations is often attributed to the former’s avoidance of vanishing/exploding gradient. The reduction in Betti numbers is significantly faster for ReLU activation than for hyperbolic tangent activation as the former

defines nonhomeomorphic maps that change topology, whereas the latter defines homeomorphic maps that preserve topology.

In sum, a deep neural network operates by gradually transforming, through its layers, a topologically complicated dataset in the input space into a topologically simple one that is linearly separable in the output space.

As regards autoencoders and deep learning, these learn the manifold structure and construct a parametric representation of it within the latent space of the autoencoder [27].

Medical Examples

The advantages of deep neural networks cited above have led to a variety of successful medical applications [8]. Health care applications of deep learning range from one-dimensional biosignal analysis [11] and the prediction of medical events, such as seizures [22] and cardiac arrests [23], to computer-aided detection [40] and diagnosis [7, 21] supporting clinical decision-making. What follows is far from comprehensive, but nevertheless important, set of examples.

Medical Imaging

Given their impressive capability at processing complex images, CNNs have become the central tool for computational medical imaging (Fig. 10). Deep learning has been applied in radiotherapy

[30], in PET-MRI attenuation correction [28], in radiomics [24], and in neurosurgical imaging [19], among others.

Genomics and Epigenomics

Machine learning methods have been widely applied to big data analysis in genomics and epigenomics research. Although accuracy and efficiency are common goals in many modeling tasks, model interpretability is especially important to these studies toward understanding the underlying molecular and cellular mechanisms. Deep neural networks have recently gained popularity in various types of genomic and epigenomic studies due to their capabilities in utilizing large-scale high-throughput bioinformatics data and achieving high accuracy in predictions and classifications.

Deep learning has been applied to popular bioinformatics problems including DNA/RNA binding sequence motif identification [36], gene expression prediction [46], epigenetic problems such as chromatin accessibility, interaction and DNA methylation predictions [9, 41], as well as various directions in noncoding RNA (ncRNA) studies [29].

Natural Language Processing

When a recurrent neural network (RNN) is unfolded over time, it can be seen to be a deep network [13], and RNNs are effective at

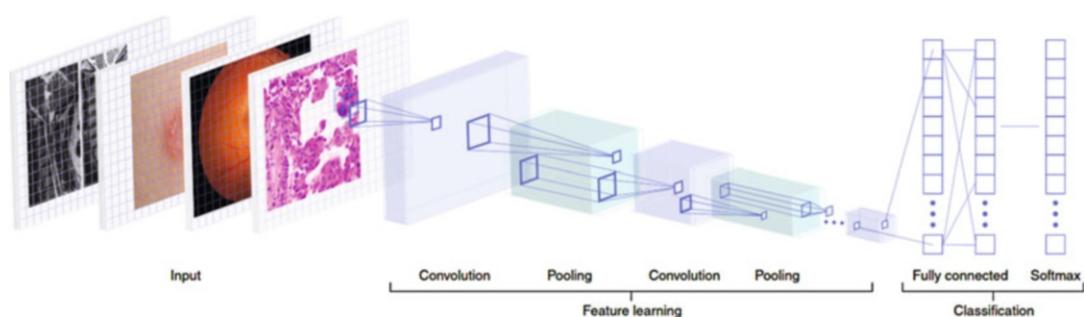


Fig. 10 Use of CNNs for medical imaging. A range of inputs can be used. Shown here are images from radiology, dermatology, ophthalmology, and histopathology [8]

processing sequential inputs such as language, speech, and time series data.

One application for RNNs in medicine is the processing of electronic patient records (EPRs). The EPR of a large medical organization can capture the medical transactions of over 10 million patients throughout the course of a decade, and a single hospitalization alone typically generates about 150,000 pieces of data [8]. One possible pipeline for building deep learning systems for EPRs is as follows [37]. Raw data are first aggregated across institutions in order to ensure that a generalizable system is built. The data are then standardized and parsed temporally and across patients, which makes them suitable for deep learning training leading to patient-specific care.

Conclusion

The topological approach to deep learning provides an eloquent way of not only understanding the nature of deep learning but also, hopefully, to facilitate the design of deep networks in a more principled manner.

It would be an understatement to say that deep learning has played a significant role in modern medicine. From medical imaging to the omics domain, progress has been impressive; however, the interpretation of deep omic-type network models toward an understanding of underlying biological mechanisms is still far from reach.

Deep learning in medical data analysis is here to stay. Even though there are many challenges associated to the introduction of deep learning in clinical settings, the methods produce results that are too valuable to discard.

References

1. Armstrong M. Basic topology. New York: Springer; 1983.
2. Barile M, Weisstein E. Betti number. 2002. <https://mathworld.wolfram.com/BettiNumber.html>. Accessed online: 2 Mar 2021.
3. Baxt W. Application of artificial neural networks to clinical medicine. Lancet. 1995;346:1135–8.
4. Bishop C. Neural networks for pattern recognition. Oxford: Clarendon Press; 1995.
5. Bishop C. Pattern recognition and machine learning. New York: Springer; 2006.
6. Cybenko G. Approximation by superpositions of a sigmoidal function. Math Control Signals Syst. 1989;2(4):303–14.
7. Dias R, Torkamani A. Artificial intelligence in clinical and genomic diagnostics. Genome Med. 2019;11:70.
8. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, ... Dean J. A guide to deep learning in healthcare. Nat Med. 2019;25:24–9.
9. Farré P, Heurtault A, Cuvier O, Embery E. Dense neural networks for predicting chromatin conformation. BMC Bioinform. 2018;19(1):372.
10. Fefferman C, Mitter S, Narayanan H. Testing the manifold hypothesis. J Am Math Soc. 2016;29(4):983–1049.
11. Ganapathy N, Swaminathan R, Deserno T. Deep learning on 1-D biosignals: a taxonomy-based survey. Yearb Med Inform. 2018;27:98–109.
12. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, MA: MIT Press; 2016.
13. Goyal P, Pandey S, Jain K. Deep learning for natural language processing. Berkeley: Apress; 2018. p. 119–68.
14. Hajij M, Istvan K. A topological framework for deep learning. arXiv, 2008.13697; 2021.
15. Hand D. Discrimination and classification. Chichester: Wiley; 1981.
16. Hochreiter S. Untersuchungen zu dynamischen neuronalen Netzen (Diplom Thesis). Josef Hochreiter Institute for Computer Science. Technical University, Munich; 1991.
17. Hornik K. Approximation capabilities of multilayer feedforward networks. Neural Netw. 1991;4(2):251–7.
18. Hubel D, Wiesel T. Receptive fields of single neurones in the cat's striate cortex. J Physiol. 1959;148(3): 574–91.
19. Izadyazdanabadi M, Belykh E, Mooney M, Eschbacher J, Nakaji P, Yang Y, Preul M. Prospects for theranostics in neurosurgical imaging: empowering confocal laser endomicroscopy diagnostics via deep learning. Front Oncol. 2018;8:240.
20. Jackson P. Introduction to expert systems. 3rd ed. Harlow: Addison Wesley; 1999.
21. Kermany D, Goldbaum M, Cai W, Valentim C, Liang H, Baxter S. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018;172:1122–31.
22. Kuhlmann L, Lehnertz K, Richardson M, Schelter B, Zaveri H. Seizure prediction – ready for a new era. Nat Rev Neurol. 2018;14:618–30.
23. Kwon J-M, Lee Y, Lee Y, Lee S, Park J. An algorithm based on deep learning for predicting in-hospital cardiac arrest. J Am Heart Assoc. 2018;7(13):118.008678.
24. Lao J, Chen Y, Li Z-C, Li Q, Zhang J, Liu J. A deep learning-based radiomics model for prediction of

- survival in glioblastoma multiforme. *Sci Rep Nature.* 2017;7:10353.
- 25. LeCun Y. Une procedure d'apprentissage pour reseau a seuil assymetrique. *Cognitiva.* 1985;85:599–604.
 - 26. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521:436–44.
 - 27. Lei N, Luo Z, Yau S-T, Gu DX. Geometric understanding of deep learning. 2018. arXiv, 1805.10451.
 - 28. Liu F, Jang H, Kijowski R, Bradshaw T, McMillan A. Deep learning MR imaging-based attenuation correction for PET/MR imaging. *Radiology.* 2018;286:676–84.
 - 29. Manzanarez-Ozuna E, Flores D-L, Gutierrez-López E, Cervantes D, Juárez P. Model based on GA and DNN for prediction of mRNA-smad7 expression regulated by miRNAs in breast cancer. *Theor Biol Med Model.* 2018;15(1):24.
 - 30. Meyer P, Noblet V, Mazzara C, Lallement A. Survey on deep learning for radiotherapy. *Comput Biol Med.* 2018;98:126–46.
 - 31. Mhaskar H, Liao Q, Poggio T. When and why are deep networks better than shallow ones? In: Singh S, Markovitch S, editors. *AAAI'17: proceedings of the thirty-first AAAI conference on artificial intelligence.* AAAI Press; 2017. p. 2343–9.
 - 32. Minsky M, Papert S. *Perceptrons.* Cambridge: MIT Press; 1969.
 - 33. Nair V, Hinton G. Rectified linear units improve restricted Boltzmann machines. In: Fürnkranz J, Joachims T, editors. *ICML'10: proceedings of the 27th international conference on international conference on machine learning.* Madison: Omnipage; 2010. p. 807–14.
 - 34. Naitzat G, Zhitnikov A, Lim L-H. Topology of deep neural networks. *J Mach Learn Res.* 2020;21:1–40.
 - 35. Olah C. Neural networks, manifolds, and topology. 2014. <https://colah.github.io/posts/2014-03-NN-Manifolds-Topology/>. Accessed online: 3 Mar 2018.
 - 36. Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.* 2016;44(11):e107.
 - 37. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, ... Dean J. Scalable and accurate deep learning with electronic healthrecords. *npj Digit Med.* 2018;1:18.
 - 38. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev.* 1958;65:386–408.
 - 39. Rumelhart D, Hinton G, Williams R. Chapter 8: Learning internal representations by error propagation. In: Rumelhart D, McClelland J, editors. *Parallel distributed processing.* Cambridge, MA: MIT Press; 1986.
 - 40. Shin H-C, Roth H, Gao M, Lu L, Xu Z, Nogues I. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging.* 2016;35:1285–98.
 - 41. Singh S, Yang Y, Póczos B, Ma J. Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. *Quant Biol.* 2019;7(2):122–37.
 - 42. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res.* 2014;15(56):1929–58.
 - 43. Tarassenko L. *A guide to neural computing applications.* London: Arnold; 1998.
 - 44. Welbank M. A review of knowledge acquisition techniques for expert systems (Memorandum No. R19/022/83). British Telecom Research Laboratories, Martlesham Heath, Ipswich IP5 7RE, UK: British Telecommunications; 1983.
 - 45. Werbos P. *Beyond regression: New tools for prediction and analysis in the behavioral sciences (PhD Thesis).* Cambridge, MA: Harvard University; 1974
 - 46. Zeng W, Wang Y, Jiang R. Integrating distal and proximal information to predict gene expression via a densely connected convolutional neural network. *Bioinformatics.* 2019;36:496503.



Suvrankar Datta

Contents

Introduction to Interventional Radiology	460
AI in Medical Imaging: A Primer	460
Data in Medical Imaging	463
AI in Diagnostic Radiology: Brief Overview	463
AI in Interventional Radiology: Unique Challenges	464
AI in Interventional Radiology: Opportunities	465
The Ideal Interventional Radiology Suite	466
Scope of AI in Interventional Radiology	466
AI in IR: Decision Support	466
AI in IR: Triaging of Patients	467
AI in IR: Prevention of Errors	467
AI in IR: Periprocedural Support	468
AI in IR: Patient Monitoring and Procedural Support	468
AI in IR: Prognostication and Outcome Prediction	469
AI in IR: Image Acquisition and Processing	469
AI in IR: Residency and Fellowship Training	469
Can AI Replace Diagnostic or Interventional Radiologists?	470
References	471

Abstract

Artificial Intelligence applications have recently demonstrated high diagnostic accuracy and increased workflow efficiency in

radiology. Machine learning models, particularly deep learning, can perform complex tasks in medical imaging, especially when they are trained with a large amount of high-quality data. For effective realization of the potential of artificial intelligence in interventional radiology, some unique challenges involving data storage, interoperability, adoption of standards, and conflict of interest between physicians and developers need to be addressed. With

S. Datta (✉)
Department of Radiodiagnosis and Interventional Radiology, All India Institute of Medical Sciences (AIIMS), New Delhi, India

immense innovation and technological breakthrough, the scope of interventional radiology is continuously increasing in both width and breadth, and so are the opportunities of AI to revolutionize the sector. Artificial intelligence can complement the efforts of the interventional radiologist through decision support, triaging and screening of patients, prevention of error, procedural and periprocedural support, patient monitoring, prognostication of diseases, outcome prediction, image acquisition, image processing, etc. In conjunction with augmented reality systems, it can also help in improving procedural skills of interventional radiology residents and fellows through superior simulation training. Artificial intelligence has a tremendous potential to boost the productivity of radiologists. However, they are unlikely to replace them as there are significant apprehensions regarding the legal accountability, transparency, fairness, equality, bias, or potential misuse by an artificial intelligence system which prevent any independent action or clinical application without the oversight of an expert radiologist.

Keywords

Radiology · Interventional radiology · IR · Artificial intelligence · AI · Machine learning · Deep learning · Medical imaging

Introduction to Interventional Radiology

Modern radiology grossly comprises two subspecialties: diagnostic radiology (DR) and interventional radiology (IR). The IR subspecialty started off as an attempt by the radiologists and physicians to use the diagnostic tools available in radiology in the management of the patient. Perhaps the most important breakthrough in interventional radiology and angiography was the introduction and popularization of percutaneous catheter replacement of a needle or a trocar by a novel technique – the Seldinger technique [1]. Since then, IR, as a subspecialty,

has gradually evolved with hundreds of innovations and improvements in the procedures and techniques which have led to significant improvement in patient outcomes [2].

While IR essentially started off as an offshoot of DR, current IR practice is very different. Interventional radiologists today are much more involved in providing holistic care for their patients. Patient evaluation, including taking a clinical history, physical examination, as well as understanding the complete laboratory workup and imaging findings prior to the interventional procedures, can guide the interventional radiologist to personalize and decide the most appropriate management plan for his patient and achieve superior treatment outcomes [3].

AI in Medical Imaging: A Primer

The use of AI in medical image analysis has garnered increased attention since the last few years, and there has been a significant rise in the curiosity and interest of radiologists and non-radiologists alike. There have been a lot of recent advancements in the application of AI in the medical imaging sector along with recent demonstrations of higher accuracy in diagnosis and increased efficiency in clinical care. Radiologists are trained during residency to visually assess medical images by scrolling through different views and sections and write a structured report to localize and characterize the findings. These reports are often subjective and vary with experience. AI can recognize complex patterns and associations from imaging data and can help in automated detection, localization, as well as quantification of lesions, wherever possible. AI-integrated workflows can help reduce workload, time to diagnosis, and errors in reporting [4] (Fig. 1).

AI in health care primarily involves the application of machine learning (ML) techniques. ML gives computers the ability to learn without being specifically programmed. It works on the principle of “backpropagation.” Forward propagation or inference is when data goes into the neural network and makes a prediction. Backpropagation is

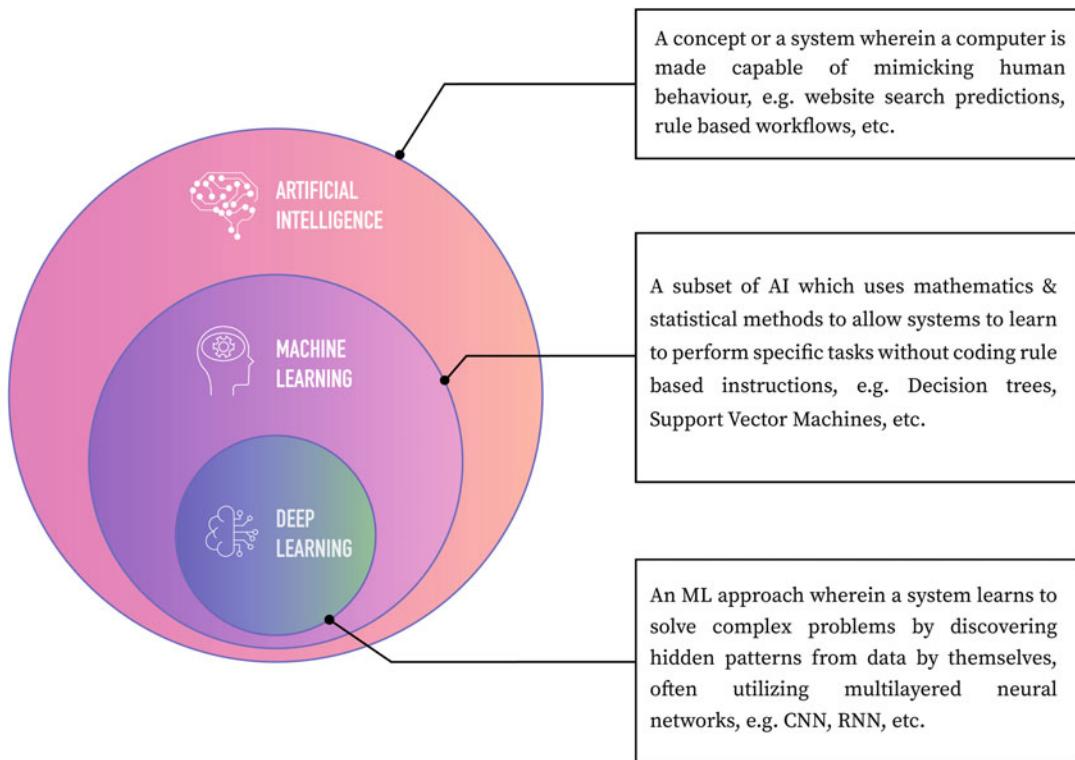


Fig. 1 Artificial intelligence (AI), machine learning (ML), and deep learning (DL)

the process of adjusting the weights by looking at the difference between prediction and the actual result, by computing how much each weight in the preceding layer needs to be changed to bring the expected outcome closer to reality, i.e., bringing the error rate close to 0. During model training, its weights are thus fine-tuned to increase the prediction accuracy. If enough high-quality data is fed to the computers, ML allows them to create algorithms for solving a particular task with high accuracy [5] (Fig. 2).

There are broadly two types of AI methods that are in wide use today – traditional methods and deep learning. Traditional AI uses pre-identified and engineered features (defined in terms of mathematical equations) and focuses on discovering relationships and confidence intervals between inputs and outcomes. These features, though perceived to be discriminative, rely on human expert definition and do not always represent the best approach for solving a particular problem. However, the traditional methods have reached their

maximum performance. Moreover, extracting and selecting classification features take up a significant amount of time and effort [6].

Compared to this, ML methods aim to reach high prediction accuracy without emphasizing on the need for interpretation. However, in health care, interpretability and explainability of any algorithm – as to understanding how and why a program is coming to its conclusions – are of paramount importance. With recent advances in ML, and the ability of the newer models to show us feature/saliency maps which highlight the region of interest (ROI) which led the AI algorithm to make a prediction, the “black box” problem of AI is an area which is being actively explored and addressed so that clinical applications of AI become more popular.

During development of an ML model, it is provided training data and tuned to make accurate predictions for the training data by an optimization algorithm. The main goals of the models are to gain insights from the training data and

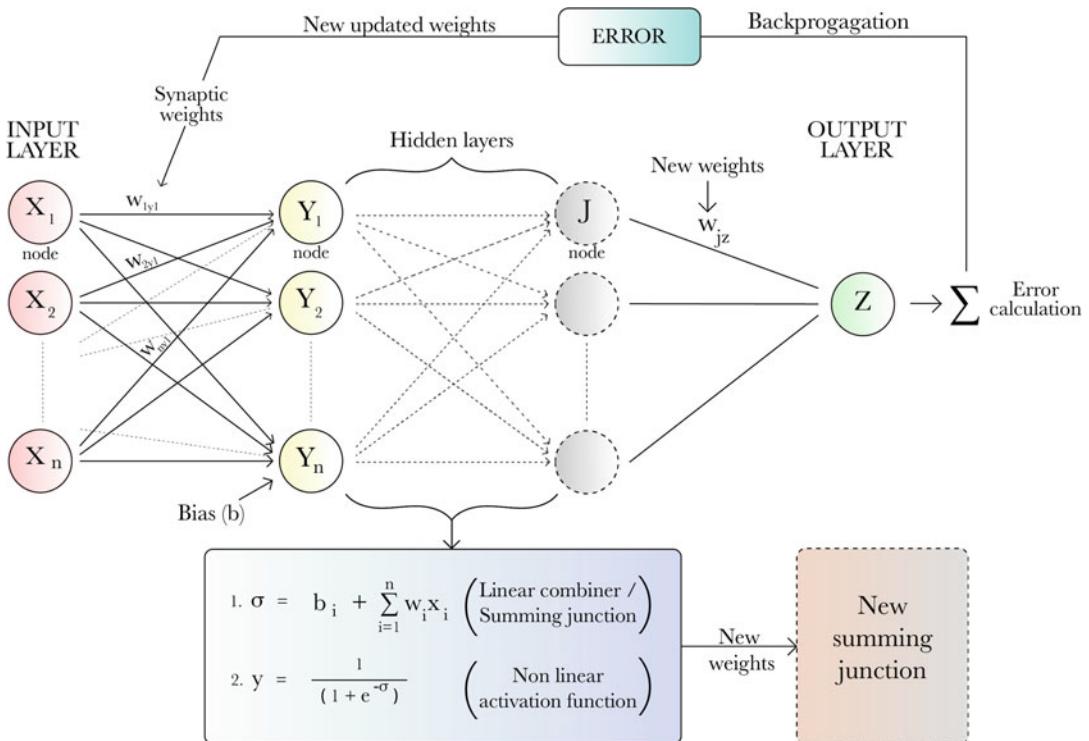


Fig. 2 An artificial neural network (ANN) and the principle of backpropagation

generalize their learned insights to make correct predictions for new, previously unseen data. The generalization ability of an ML model is typically estimated using a separate data set – the validation set and the same is used as feedback for further tuning of the model. After multiple iterations of training and tuning, the final model is evaluated on a test dataset to simulate how the model will perform when faced with new, unseen data.

Deep learning (DL) is a subset of machine learning, and, while it has similar functions, its capabilities are different. DL uses multiple layers of artificial neural networks, with the capability of learning complex tasks. While most ML models may perform fairly well with structured and organized data, DL models have the capacity to process medical images and other high-dimensional and less structured data with excellent results [5].

Without explicit feature selection, DL algorithms have the ability to learn complex interrelationships and associations by navigating the data space which provides them superior

problem-solving capabilities. While various DL techniques have been explored to address different medical imaging tasks, convolutional neural networks (CNNs) are the most well-known DL architecture today. A CNN comprises a series of layers that can map image inputs to desired end points while learning increasingly higher-level imaging features. Starting from an input image, there are multiple hidden layers within CNNs (including convolution and pooling operations) for extracting feature maps and performing feature aggregation. These hidden layers are followed by fully connected layers providing high-level reasoning before an output layer produces predictions. CNNs are often trained end to end with labeled data for supervised learning. Compared to traditional AI algorithms, CNNs can learn and select the features by itself automatically and effectively. However, for the better adoption of CNNs in clinical applications, especially in medical imaging, visualization of ROI is essential for the interpretability of its results [6].

Data in Medical Imaging

Data in medical imaging has tremendous potential to complement efforts of healthcare providers and radiologists in improving patient outcomes. Radiology departments around the world currently have huge databases of images and text. They are often rich in data but poor in information content which can be effectively utilized. In IR, a vast majority of the data are stored as video files which take up significant storage space. “Big Data” in radiology represents an opportunity to utilize reservoirs of information to make medical imaging more valuable to patients, e.g., by providing clinical decision support to radiologists through access to and analysis of information that is otherwise often inaccessible, inefficient, or difficult to integrate in real time for the consulting physician.

One of the key issues which is a primary concern regarding the utilization of medical data is interoperability. Medical information in the form of electronic health records (EHRs) is collected and stored with great variation across different institutes and organizations. Most software programs utilize and share information in proprietary formats, e.g., free flowing text notes for a particular interventional procedure, unstructured or semi-structured radiology reports, etc., and much of the data is often missing, inaccurate, and non-interpretable across other platforms. This poses significant challenges during the analysis of data consolidated from inconsistent and often different sources [7].

Despite advancements in interoperability, such as Consolidated Clinical Document Architecture (C-CDA) and Fast Healthcare Interoperability Resources (FHIR), there are multiple reasons which have hindered the widespread and interinstitutional exchange of medical imaging data, ranging from issues involving selection and adoption of technical and image standards to governance, security, and privacy [8].

Big Data powered by AI and statistical modeling can predict the pre-test probability of a disease for a patient based on his or her profile as extracted from the electronic health record, in real time, and

help the radiologist or clinician in interpreting the reports of a diagnostic imaging study, e.g., during analysis of a pulmonary nodule incidentally detected in a patient or for taking management decisions by determining and comparing the likelihood of success of a particular IR procedure based on patient profile, e.g., decision regarding splenic artery embolization vs. splenectomy in a Grade III splenic trauma.

For complete realization of the prospects of Big Data in medical imaging and radiology, increased information sharing among organizations is a prerequisite. Medical data, especially medical image/video sharing, is very complex and requires patients, radiologists, clinical departments, administrators, and institutions to work together toward democratization of anonymized healthcare information and adoption of safe data-sharing standards. Strategies for creating image data sets from multiple large cohorts for integration and comparison, addressing security and accessibility concerns for researchers and clinicians, and the standardization of information are crucial for the effective utilization of the large volume of medical imaging data available worldwide. Widespread global collaborative efforts sound ambitious but are indeed necessary to unleash the true potential of data in both DR and IR.

AI in Diagnostic Radiology: Brief Overview

One of the most promising clinical applications of AI is in diagnostic imaging, and it is indeed on the road to revolutionize DR as we know of today. Currently, increased attention is being directed at establishing and fine-tuning its performance to facilitate detection and characterization of a multitude of clinical conditions [9].

Improvement in imaging techniques, introduction of newer technologies, and the wider availability of imaging has resulted in increased reporting workload of radiologists over the past decades. This increase in demand is particularly prevalent in more time-consuming imaging such

as computed tomography (CT) and magnetic resonance imaging (MRI), with the additional challenge of completing the reports in a timely manner. Incorporation of AI into the radiologists' workflow can increase reporting efficiency as well as improved patient outcomes.

AI helps in automation of lesion detection, image segmentation, lesion quantification, standardization of radiology reports, or for comparison of scans with previous imaging. Follow-up studies for monitoring treatment response or for disease relapse often take up significant time which can be cut short by AI-assisted lesion measurement and characterization. These mundane and laborious tasks may benefit from automation using AI and allow radiologists to focus on other complex and cognitive tasks. Natural language processing (NLP) of radiology reports is also an area which is being actively explored. DL algorithms can insert recommendations for the clinician at the end of the report whenever critical findings are detected on a scan [10].

AI can also help in the scheduling of orders, triaging or screening of patients for scans or procedural safety, examination protocoling, faster image acquisition, image management and archiving, image registration, quality analytics, dose estimation, and integration of reports or imaging data with EHRs. These can help in improving efficiencies and patient outcomes by reducing missed radiology appointments, decreasing adverse events and mistakes, decreasing imaging time, reducing unnecessary imaging and radiation exposure, optimization of sequences, prioritization of worklists, improved diagnostic interpretation, prognosis estimation, and automatic quantification of lesions/bleeds.

However, despite the tremendous advancements and potential to disrupt the DR sector, a few key challenges still need to be addressed. Collection of high-quality annotated ground truth data, development of generalizable and diagnostically accurate techniques, and workflow integration are the key challenges facing adoption of machine learning in radiology practice. Overfitting is also an important challenge in complex models and tasks. ML develops high-dimensional functions that often cannot be explained. This makes interpretability difficult, thereby hindering

its acceptance in health care where identifying the underlying logic is important. Visual saliency maps are being increasingly incorporated in AI solutions which could highlight areas within images that have grabbed the attention to perform a classification task. The saliency maps could provide a certain extent of explainability for the ML models. Widespread application of ML algorithms in DR is expected to improve radiology workflow, increase productivity, and improve patient outcomes [11].

AI in Interventional Radiology: Unique Challenges

Apart from the challenges common to both DR and IR, there are indeed a few unique issues in the latter which need to be addressed for the successful utilization of AI in interventional procedures to improve workflow efficiency and outcomes. IR images/videos are essentially stacks of image data or sequential scans, which have unique spatiotemporal associations. Such data needs to be stored and retrieved in an organized and sequential manner when an attempt is made to create datasets for ML model development. These datasets occupy a significant amount of storage space.

One major limitation is the significantly less availability of expert interventional radiologists globally. Creation of high-quality datasets often require establishment of ground truth by means of annotations by experts. While there are a large number of diagnostic radiologists who are interested in collaborations for the development of AI tools or in the incorporation of AI in their workflows, the interest among interventional radiologists is comparatively less.

Conflict of interest between interventional radiologists and AI developers is an important issue which needs to be addressed. While many AI developers are enthusiastic about creating a "do-it-all" product which can replace radiologists, they need a lot of support with respect to domain expertise from the radiologists while developing their product. Radiologists are in turn often wary about their own careers and the potential of AI to impact their DR or IR practice. IR is significantly

different from DR as the former involves procedures that alter patient management immediately, and there may be significant ethical concerns when intraprocedural decisions based on recommendations of an “inexplicable” ML model are to be factored in. This may have major medicolegal implications and is a potential barrier to the effective incorporation of AI in IR.

Interventional procedures, empowered by the fast-paced introduction of innovations in the sector, are dynamic, and with rapidly changing practice patterns and new evidence being generated, best practice guidelines for IR procedures are being continuously updated. This limits the pool of homogenous and retrospective data available for training ML algorithms [12].

AI in Interventional Radiology: Opportunities

Before discussing the detailed scope of AI in improvement of IR, it is important to know the different procedures that an interventional radiologist deals with, in his clinical practice. With breakthroughs in technological innovation, the opportunities in IR continue to expand both in width and depth. IR procedures can be broadly divided into vascular and non-vascular interventions [13].

Vascular interventions may be aimed at:

- Improving the vessel lumen by balloon angioplasty, thrombectomy, stents, or stent-graft placements
- Decreasing the blood flow by embolization or endovascular ablation
- Implantation of devices, e.g., filters and transjugular intrahepatic portosystemic shunt (TIPS)
- Intravascular foreign body retrieval
- Transvascular biopsy

Important vascular conditions where IR plays an important role in management:

- Superior vena cava (SVC) syndrome
- Pulmonary embolism
- Pulmonary artery tumors, aneurysms, and pseudoaneurysms

- Aortic aneurysms
- Uterine fibroids
- Benign prostatic hyperplasia
- Mesenteric ischemia
- Gastrointestinal (GI) bleeding
- Solid organ (e.g., liver or spleen) injury
- GI neoplasms
- Visceral artery aneurysms and pseudoaneurysms
- Renal artery occlusive disease
- Renal neoplasms, e.g., angiomyoma
- Arteriovenous malformations (AVMs) and fistulas
- Pelvic congestion syndrome
- Hepatic vein obstruction/Budd-Chiari syndrome
- Venous thromboses and insufficiencies

Important non-vascular IR procedures include:

- Image-guided percutaneous biopsy of intra-abdominal solid organs
- Lung and mediastinal biopsy
- Diagnostic fluid aspiration or drainage procedures of intra-abdominal abscesses
- Thoracocentesis and empyema drainage
- Percutaneous gastrostomy
- Percutaneous gastrojejunostomy and gastrostomy tube conversions
- Stent placement for malignant GI/biliary strictures
- Percutaneous biliary drainage
- Percutaneous extraction of common bile duct (CBD) stones
- Gall bladder aspiration/ablation
- Percutaneous nephrostomy, lithotripsy, and nephrolithotomy
- Balloon dilatation of ureteric strictures
- Suprapubic cystostomy
- Ultrasound, MRI, or stereotactic-guided breast biopsy
- Breast lesion excision systems
- Percutaneous epidural and nerve root block
- Bipolar radiofrequency nucleoplasty
- Percutaneous ablation of bone tumors
- Percutaneous vertebroplasty and disc decompression
- Image-guided ablation of renal, lung, and liver tumors

The Ideal Interventional Radiology Suite

It is a good idea to understand how an ideal interventional radiology suite (henceforth referred to as the “IR suite”) looks like for a better understanding of the possibilities for innovation and integration of AI into the workflow. The IR suite is the actual procedural room and is a component of the broader “IR area” which also includes the adjacent control rooms, corridors, additional storage spaces, and picture archiving and communication system (PACS) workstation. The “periprocedural area” refers to the preparatory (pre-procedure) and recovery (post-procedure) patient areas. The IR area usually should maintain some basic recommended standards to optimize patient flow, efficiency, and procedural safety. The following is a depiction of an ideal IR area compliant with the recommendations by the Society of Interventional Radiology (SIR) [14].

The periprocedural area is in the vicinity of the IR suite which ensures maximum efficiency and workflow. It includes adequate space for privacy during discussions (e.g., for obtaining informed consent) and physical examination. Three to four preparatory and recovery rooms per IR suite with contingency plans to accommodate unanticipated urgent patient referrals should be adequate. Programs providing IR fellowship/residency training should also have adequate facilities and space for the education and training including access to computers. The periprocedural area must have adequate electrical, oxygen, suction, anesthesia, and emergency equipment and services essential for basic resuscitation. Access to advanced resuscitation equipment, medications and fluids, and nursing facilities is also essential. The scrubbing area with sinks is immediately outside the IR suite (similar to an operation theater).

SIR recommends 650 square feet for an IR suite which along with the IR area must be sufficiently big to accommodate the personnel and equipment required for a safe procedural environment and manage emergencies. For an IR/operating room hybrid suite, at least 800 square feet is recommended.

IR suite should have a permanent, mounted, C-arm fluoroscopy unit and also have access to an

ultrasound machine and a CT scanner. In low-volume setups, a portable C-arm unit may be used, if safe. In setups with a high load of CT-guided interventions, the IR suite should have a CT scanner. This improves patient access and workflow efficiency. Equipment for advanced cardiac life support and power injectors must be accessible. Contingency plans should be in place to address procedural complications, intraprocedural conversions, and all possible emergencies.

Apart from the main IR suite, space for controlling the imaging equipment and a dedicated image viewing, processing, analysis, and reporting area for the interventional radiologists should be in close proximity to the IR suite. PACS and EHRs should be readily accessible from the radiologists’ workspace [14, 15] (Fig. 3).

Scope of AI in Interventional Radiology

AI has the potential to complement the interventional radiologist at multiple points of time during the entire course of a procedure – right from the selection of a patient or candidate till the follow-up of a patient following discharge from the recovery room.

AI can help in IR in the following broad areas (each of which will be discussed elaborately in following subsections):

- Decision support
- Triaging of patients
- Prevention of error
- Periprocedural support
- Patient monitoring and procedural support
- Prognostication and outcome prediction
- Image acquisition and processing
- Residency and fellowship training

AI in IR: Decision Support

When a patient is referred to the interventional radiologist, the physician has a lot of decisions to make regarding the procedure which has the best possible outcomes for the patient. AI has a tremendous potential to help the interventionalist

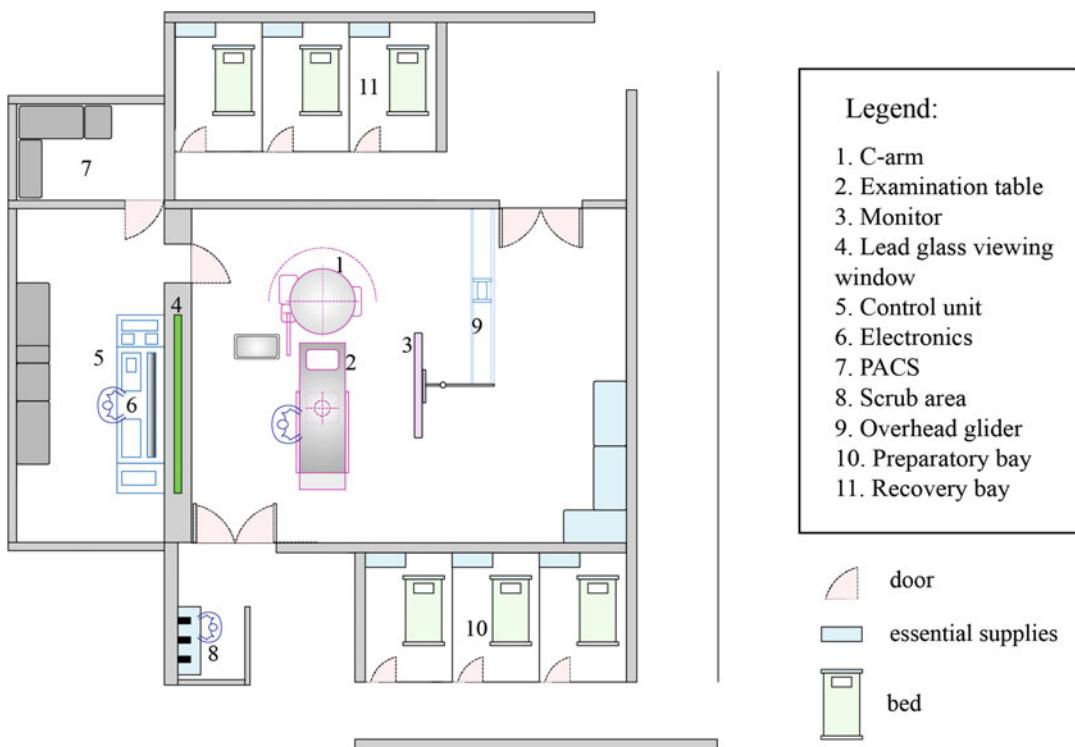


Fig. 3 Floor plan or architectural representation of an interventional radiology (IR) suite

in earlier diagnosis of diseases and also help in making treatment choices. Most of the available and widely used staging/grading systems which guide management of conditions like cancers do not provide personalized treatment choices [12]. AI-powered decision support tools can factor individual patient characteristics based on demographics and clinical and imaging profile and help in arriving at the best IR procedure (if it is at all required). This can significantly reduce treatment costs and also decrease unnecessary procedures which do not alter outcomes (e.g., life expectancy) significantly. An AI-integrated system can extract a patient's physical characteristics from prior scans and help the radiologist in choosing the equipment for a procedure, e.g., appropriate size of instruments like catheters, micro-puncture systems or needles, etc.

AI in IR: Triaging of Patients

AI can help in triaging of patients and help radiologists prioritize candidates in need for a particular interventional procedure (e.g., splenic artery

embolization in splenic injury), particularly in institutes with a high volume of procedures. Triaging of patients based on their radiographic images has already been shown to be successful and fairly accurate [16]. AI can help in analyzing the CT scans done in a trauma/emergency setup, detect suspected lesions/bleeds, and prioritize the scans in the radiologists' worklist. This can help in alerting the IR physician of an emergent procedure (e.g., embolization to stop a life-threatening bleed or for thrombectomy in a stroke patient) and help him modify the ongoing procedures and upcoming scheduled procedures to address the urgent case based on severity and clinical judgment.

AI in IR: Prevention of Errors

To err is human and making errors in medicine is inevitable. Today, we undoubtedly have a better understanding of medical error than we had before, but most organizations fail to respond to adverse events in a way that prevents the recurrence of such errors. Medical errors most likely

occur during routine procedures due to unanticipated patient complexity or system failures. While checklists are an effective way to prevent error, there are barriers to adoption of checklists in IR – lack of awareness, effective leadership, staff attitude, etc. A majority of errors in IR are preventable – ranging from ineffective procedure planning and communication error to equipment difficulties [17].

Incorporating AI into the workflow can automate mundane tasks like procedural planning and preprocedural checklisting (e.g., contrast allergy or previous adverse events from EHRs). In the future, AI-integrated tracking systems in the IR suite will be able to prevent errors of both commission (e.g., injection of wrong solutions) and omission (e.g., missed history of contrast allergy). Intelligent tracking systems can help identify missed foreign bodies and malpositioned/unsecured devices and prevent wrong site procedures. AI can also remind about periodic functionality checks of essential equipment [18].

AI in IR: Periprocedural Support

AI can give preprocedural support by automating procedural consent and curated history taking, tracking physical characteristics like weight, meals, and activity with the data being continuously fed into their EHRs. Automated analysis of all preprocedural data would provide a more personalized and precise risk stratification score which can play a significant role in the procedural planning and determine valuable predictors for post-procedural care and outcome.

Integration and analysis of periprocedural and intraprocedural data can further help in post-procedural discharge/transfer planning, prediction of complications, long-term outcome prediction, and curated follow-up appointment scheduling. Post transfer, post-procedural monitoring of vitals and other relevant data could continue to be integrated with EHRs to predict potential adverse events and set off an alarm. This will lead to achievement of truly personalized and patient-centered care [19].

AI in IR: Patient Monitoring and Procedural Support

Similar to surgeons, interventional radiologists usually make decisions through a three-step process that includes situational assessment, taking an action and reevaluation of the action's consequences [20]. AI could help the radiologist in better assessment of an intraprocedural situation (e.g., empowerment of the interventionalist with preoperative data and intraoperative monitoring of vitals), the types of actions that are taken (e.g., through decision support tools), and the process of reevaluating the impact of an action (e.g., through predictive analytics based on the real time vitals monitoring).

Particularly in IR procedures, AI-assisted analysis of intraprocedural imaging data, e.g., data generated from fluoroscopy, digital subtraction angiography (DSA), or CT has tremendous potential to guide the interventionalist more accurately for the next steps. For example, during CT-guided percutaneous drain insertion for an intra-abdominal abscess drainage, an intelligent AI-assisted instrument/hand tracking system can dictate how much more insertion of the catheter is required and in which direction, for the catheter tip to reach the particular region of interest inside the abscess for adequate drainage. Similar assistance can be provided for negotiating vessels during angiography by analysis of DSA images. This can lead to significant reduction in exposure of both the patient and the radiologist to harmful radiation.

In the future, AI will augment intraprocedural decision-making by the interventional radiologist based on real-time analysis of procedural progress that utilizes integrated data from EHRs, procedural video, vital signs, instrument/hand tracking, and electrosurgical energy usage. Monitoring of such intraprocedural data can facilitate real-time prediction and avoidance of significant adverse events (e.g., accidental exposure of staff or healthcare workers not wearing lead aprons by intelligent logic gates/switches which prevent exposure when any staff inside an IR suite is not wearing protective equipment) [21].

Another interesting application of AI would be in decreasing the documentation workload of the interventional radiologist by real-time update of procedure notes in the patient EHR either by voice-assisted intraprocedural dictations or by more intelligent tracking systems which can write structured procedural notes by tracking intraprocedural events. This would allow the IR physician to focus primarily on complex cognitive tasks and save a significant amount of time.

AI in IR: Prognostication and Outcome Prediction

AI has demonstrated fairly accurate results in the prognostication of diseases and prediction of post-intervention outcomes, e.g., mortality prediction in stroke and outcome prediction in arteriovenous malformations post-endovascular ablation [22]. Currently, multiple prognostic scoring systems for patient stratification are available for a single disease or condition. Clinicians are often unclear on which is most useful. Most of the systems do not factor in all patient characteristics and risk factors while predicting prognosis. AI, powered by big data, can extract useful information from patients' EHRs, identify key determinants, and help in creation of personalized and comprehensive prognostication systems. Traditional AI models in IR have shown promising results in predicting treatment responders from non-responders by preprocedural medical image analysis. These can help in the selection of the best suited IR procedure for a candidate.

AI in IR: Image Acquisition and Processing

Usually at the beginning of any IR procedure, once a baseline scan has been obtained (e.g., via DSA or CT), it is not always essential to have the same high quality/resolution of image for successive scans. Subsequent scans can be obtained at significantly lower radiation doses and optimized by AI-integrated processing systems, by

extracting key anatomical and imaging features from the prior scan. This leads to significant reduction in radiation exposure. Prior works in CT have also demonstrated the capabilities of AI in optimization of data acquisition processes by automating patient positioning and acquisition parameter settings. Post-data acquisition, AI-enabled optimization of image reconstruction, advanced reconstruction algorithms, and image denoising can help in improving image quality [23].

While most interventional procedures are done under ultrasound, fluoroscopy/DSA, and CT guidance, magnetic resonance imaging (MRI)-guided interventions like biopsies involving musculoskeletal system or breast, periradicular therapy and injections, and thermal therapies are also becoming common. Imaging sequences in interventional MRI are different from those used in diagnostic MRI to achieve faster imaging speed [24]. In the trade-off between image speed, signal-to-noise ratio, and resolution, AI can help in faster image acquisition by utilizing DL methods for reconstructing undersampled MRI data and generating high-resolution from low-resolution data [25].

AI in IR: Residency and Fellowship Training

Residency and fellowship training programs in IR are aimed at ensuring that their trainees are capable of proficiently performing interventional procedures which they may encounter in their career with a minimum prerequisite skill level necessary for successful patient outcomes. However, the case load and volume load of patients are different in different setups, and a resident during the tenure of his or her training and rotations may not get exposed to important interventional procedures. AI has a potential to support IR residency training in combination with augmented reality (AR) systems and provide trainees a platform for honing their procedural skills, especially in low-volume cases. Recent work in multiple surgical specialties has demonstrated the effectiveness

of simulation tools for procedural training. Patient-specific anatomic data which is currently obtained by AR systems from cross-sectional imaging by traditional image segmentation methods could be made easier and automated by ML models. Such simulation training can help in the improvement of skills of the next generation of interventional radiologists and indirectly improved health outcomes for the patients [12, 26].

Can AI Replace Diagnostic or Interventional Radiologists?

AI has a tremendous potential to boost the productivity of the interventional radiologist by helping them in various ways mentioned in the preceding subsections. In recent times, the belief that AI is going to replace the human radiologist altogether has gained traction. Such assumptions are due to the perception of radiology as a science where the physician is involved in the analysis of images and that advancement in AI can help build algorithms which could be more accurate and better than humans. AI will indeed replace many of the time-consuming works of radiologists, e.g., detection of rib fractures by scrolling through multiple CT sections and scanning all 24 ribs, thereby allowing radiologists to perform higher-level tasks which require cognitive skills.

Most AI algorithms in medical imaging look at very specific problems. However, a radiologist takes into account the clinical perspective by going through a patient's history and investigations and looks for a number of pathologies in the scan. Current AI models may be better in testing conditions and in detecting specific findings, but they are not yet capable of detecting all findings and integrating all information to provide curated differential diagnoses for a patient. Interventional radiologists also use their cognitive abilities, training, and experience to decide the best procedure based on these differentials. While, in the future, an AI system which works in conjunction with a highly integrated system with access and capability to extract information from standardized EHRs and PACS may be able to factor all available

information for a patient and provide a diagnosis, to assume that an AI-powered robot will be able to utilize clinical, laboratory, and radiological information of a patient to arrive at a diagnosis and subsequently select and perform an interventional procedure without the involvement of an interventional radiologist seems to be a far-fetched dream [27].

The tendency for humans to favor machine-generated decisions, ignoring contrary data or conflicting human decisions, also known as automation bias, can lead to significant errors when an operator is overly dependent on an AI system for taking decisions. While AI promises to improve health care in resource-poor setups, automation bias risks may be significant in such populations because there is no local expert/radiologist to effectively veto the decisions taken by the AI system.

AI performs relatively well with a lot of complex tasks, but it is important to note that it is not human. Accountability is a major issue with AI products, and it is essential that its results and actions be monitored by a human expert. Transparency, fairness, and equality are not, in particular, concepts which a machine understands. Human operators are responsible for these insights who must be trained enough to anticipate how changing AI models may perform incorrectly or misused and protect against an unethical outcome before one may occur [28].

Thus, even with massive technological advancements, multiple key issues and ethical dilemmas need to be addressed before AI models are permitted to take even minor decisions which can affect or alter patient management. An independent AI system which can perform all the functionalities of a human diagnostic or interventional radiologist without their oversight does not seem to be a possibility in the foreseeable future. How much ever advanced an AI system might become, they need to be operated by an expert who can oversee, detect, and prevent adverse outcomes. The decision-making and legal accountability still lie and will continue to lie with the expert operator, i.e., the radiologist [29].

Modern training programs should teach radiologists the advantages, limitations, and clinical use

of AI-based systems. Without radiology expertise, an accurate outcome and successful translation of AI systems from research to clinical practice will not be possible. Thus, the way forward for AI developers is to work together with the radiologists (both diagnostic and interventional) to achieve successful clinical translation of their products and complement them in improving patient outcomes rather than being over-ambitious about replacing them.

References

- Seldinger SI. Catheter replacement of the needle in percutaneous arteriography; a new technique. *Acta Radiol.* 1953;39(5):368–76.
- Baum RA, Baum S. Interventional radiology: a half century of innovation. *Radiology.* 2014;273(2S):S75–91.
- Taslakian B, Georges Sebaaly M, Al-Kutoubi A. Patient evaluation and preparation in vascular and interventional radiology: what every interventional radiologist should know (Part 1: Patient assessment and laboratory tests). *Cardiovasc Intervent Radiol.* 2016;39(3):325–33. <https://doi.org/10.1007/s00270-015-1228-7>.
- Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer.* 2018;18(8):500–10. <https://doi.org/10.1038/s41568-018-0016-5>.
- Meskó B, Görög M. A short guide for medical professionals in the era of artificial intelligence. *npj Digit Med.* 2020;3:126.
- Yadav SS, Jadhav SM. Deep convolutional neural network based medical image classification for disease diagnosis. *J Big Data.* 2019;6:113.
- Morris MA, Saboury B, Burkett B, Gao J, Siegel EL. Reinventing radiology: big data and the future of medical imaging. *J Thorac Imaging.* 2018;33(1):4–16.
- Persons KR, Nagels J, Carr C, Mendelson DS, Fischer B, Doyle M. Interoperability and considerations for standards-based exchange of medical images: HIMSS-SIIM Collaborative white paper. *J Digit Imaging.* 2020;33(1):6–16.
- Oren O, Gersh BJ, Bhatt DL. Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. *Lancet Digit Health.* 2020;2(9):e486–8.
- Lee LI, Kanthasamy S, Ayyalaraju RS, Ganatra R. The current state of artificial intelligence in medical imaging and nuclear medicine. *BJR| Open.* 2019;1:20190037.
- Choy G, Khalilzadeh O, Michalski M, Do S, Samir AE, Pianykh OS, Geis JR, Pandharipande PV, Brink JA, Dreyer KJ. Current applications and future impact of machine learning in radiology. *Radiology.* 2018;288(2):318–28.
- Meek RD, Lungren MP, Gichoya JW. Machine learning for the interventional radiologist. *Am J Roentgenol.* 2019;213(4):782–4.
- Kaufman JA, Lee MJ. Vascular and interventional radiology: the requisites e-book. Elsevier Health Sciences; 2013.
- Baerlocher MO, Kennedy SA, Ward TJ, Nikolic B, Bakal CW, Lewis CA, Winick AB, Niedzwiecki GA, Haskal ZJ, Matsumoto AH. Society of interventional radiology: resource and environment recommended standards for IR. *J Vasc Interv Radiol: JVIR.* 2017;28(4):513.
- Taslakian B, Ingber R, Aaltonen E, Horn J, Hickey R. Interventional radiology suite: a primer for trainees. *J Clin Med.* 2019;8(9):1347.
- Annarumma M, Withey SJ, Bakewell RJ, Pesce E, Goh V, Montana G. Automated triaging of adult chest radiographs with deep artificial neural networks. *Radiology.* 2019;291(1):196–202.
- Mafeld S, Musing EL, Conway A, Kennedy S, Oreopoulos G, Rajan D. Avoiding and managing error in interventional radiology practice: tips and tools. *Can Assoc Radiol J.* 2020;71:528–35. <https://doi.org/10.1177/0846537119899215>.
- Mafeld S, Oreopoulos G, Musing EL, Chan T, Jaber A, Rajan D. Sources of error in interventional radiology: how, why, and when. *Can Assoc Radiol J.* 2020;71:518–27. <https://doi.org/10.1177/0846537119899226>.
- Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial intelligence in surgery: promises and perils. *Ann Surg.* 2018;268(1):70.
- Flin R, Youngson G, Yule S. How do surgeons make intraoperative decisions? *BMJ Qual Saf.* 2007;16(3):235–9.
- Navarrete-Welton AJ, Hashimoto DA. Current applications of artificial intelligence for intraoperative decision support in surgery. *Front Med.* 2020;14:369–81.
- Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, Wang Y. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol.* 2017;2(4):230–43.
- McCollough CH, Leng S. Use of artificial intelligence in computed tomography dose optimisation. *Ann ICRP.* 2020;49:113–25. <https://doi.org/10.1177/0146645320940827>.
- Blanco Sequeiros R, Ojala R, Kariniemi J, Perälä J, Niinimäki J, Reinikainen H, Tervonen O. MR-guided interventional procedures: a review. *Acta Radiol.* 2005;46(6):576–86.
- Johnson PM, Recht MP, Knoll F. Improving the speed of MRI with artificial intelligence. *Semin Musculoskelet Radiol.* 2020;24(1):12. NIH Public Access.
- Mirchi N, Bissonnette V, Yilmaz R, Ledwos N, Winkler-Schwartz A, Del Maestro RF. The virtual operative assistant: an explainable artificial intelligence

- tool for simulation-based training in surgery and medicine. *PLoS One.* 2020;15(2):e0229596.
27. Chander Mohan SM. Artificial intelligence in radiology—are we treating the image or the patient? *Indian J Radiol Imag.* 2018;28(2):137.
28. Geis JR, Brady AP, Wu CC, Spencer J, Ranschaert E, Jaremko JL, Langer SG, Kitts AB, Birch J, Shields WF, van den Hoven van Genderen R. Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement. *Can Assoc Radiol J.* 2019;70(4):329–34.
29. European Society of Radiology (ESR). Impact of artificial intelligence on radiology: a EuroAIM survey among members of the European Society of Radiology. *Insights Imag.* 2019;10(1):105.



Automated Deep Learning for Medical Imaging

33

Ciara O'Byrne, Laxmi Raja, Robbert Struyven, Edward Korot,
and Pearse A. Keane

Contents

Introduction	474
Challenges of Deep Learning for Clinical Researchers	475
Highly Specialized Technical Expertise	475
Compute Resources	475
Large, Well-Curated Datasets	476
Data Protection and Privacy	476
Principles	477

C. O'Byrne
Moorfields Eye Hospital NHS Foundation Trust,
London, UK

Trinity College Dublin, Dublin, Ireland
e-mail: c.obyrne@nhs.net

L. Raja
Moorfields Eye Hospital NHS Foundation Trust,
London, UK

University College London, London, UK

R. Struyven
Moorfields Eye Hospital NHS Foundation Trust,
London, UK

E. Korot
Moorfields Eye Hospital NHS Foundation Trust,
London, UK

Byers Eye Institute, Stanford University, Palo Alto,
CA, USA

P. A. Keane (✉)
NIHR Biomedical Research Centre for Ophthalmology,
Moorfields Eye Hospital NHS Foundation Trust,
London, UK
e-mail: p.keane@ucl.ac.uk

Initial Work Utilizing Automated Deep Learning in Medicine	478
Automated Deep Learning Process	479
Limitations of Automated Deep Learning	481
General Limitations	481
Automated Deep Learning-Specific Limitations	482
Future Directions of Automated Deep Learning	483
Use Cases of Automated Deep Learning	483
Packaging and Deployment of These Models	483
Conclusion	484
References	484

Abstract

Automated deep learning is a subset of machine learning. It aims to automate the machine learning workflow allowing those with limited or no coding expertise to create deep learning algorithms. It is available on a number of commercial platforms. A number of limitations still exist for automated deep learning that clinicians must be aware of. Datasets must still be curated and labelled and data governance obstacles must be navigated. Additionally, the challenges of interpretability, generalizability, and bias still exist. Automated deep learning for medical imaging has demonstrated promising results within the clinical literature when compared against bespoke machine learning models. It has generated considerable excitement as it offers the potential to democratize artificial intelligence in healthcare. In the following chapter, we will explore the role of automated deep learning within the rapidly progressing field of clinical artificial intelligence. We will examine its challenges and limitations, the principles and process of use, and what we consider the future directions of automated deep learning to be.

Keywords

Automated deep learning · Automated machine learning · AutoML · Medical imaging · Neural architecture search · Reinforcement learning · Automated deep learning platforms

Introduction

They are a dream of researchers but perhaps a nightmare for highly skilled computer programmers: artificially intelligent machines that can build other artificially intelligent machines. New York Times, 2017 [1]

Deep learning has emerged as a form of machine learning anticipated to transform a variety of industries, with major breakthroughs in technology, government, science, and healthcare [2–4]. However, along with great promise comes an array of challenges. These are particularly relevant to the clinical research field, where lack of resources and specialized technical expertise can represent significant barriers. In 2017, the New York Times spotlighted the emerging area of automated deep learning, described as “part of a much larger effort to bring the latest and greatest A.I. techniques to a wider collection of companies and software developers” [1].

The need to increase accessibility of deep learning is well recognized, and initial efforts have included open sourcing of libraries (e.g., TensorFlow, PyTorch), the use of techniques such as transfer learning, and automation of parts of the machine learning pipeline [5]. However, these approaches still require computer expertise. The arrival of automated deep learning has been a major step forward in opening deep learning to a greater range of users. Automated deep learning is a subset of machine learning and aims to automate the machine learning workflow, enabling those with limited or no coding experience to create deep learning algorithms (Fig. 1). It is available on a number of commercial platforms including

MACHINE LEARNING: Dotted Lines represent parts of the process that can be automated (AutoML)



COMMERCIAL PLATFORMS using AUTOMATED DEEP LEARNING: Automation of Neural Architecture Search



Fig. 1 Comparison of traditional machine learning, automated machine learning, and automated deep learning

Amazon Rekognition Custom Labels (Amazon), Apple Create ML (Apple), Clarifai Train (Clarifai), Google Cloud AutoML Vision (Google), MedicMind Deep Learning Training Platform (MedicMind), and Microsoft Azure Custom Vision (Microsoft). Automated deep learning has generated considerable excitement across multiple industries. In medicine, it offers the potential of empowering clinical researchers, enabling them to much more easily explore new novel clinical applications of AI and develop new research tools.

Herein, we will investigate the role of automated deep learning within the greater evolving medical machine learning landscape. We will describe an overview of the challenges and limitations, the technical principles, and the process for users. Finally, we will discuss what we envisage the future directions of automated deep learning to be.

AlexNet in the ImageNet Challenge in 2012, major technology companies such as Google and Facebook began recruiting researchers with expertise in deep learning [6] significantly reducing the availability of deep learning experts for clinical academic research. While the basic concepts of deep learning are not difficult to grasp, expertise in its development and application requires significant mathematical knowledge beyond that normally attained by computer science and engineering graduates. As a result, the competition for graduates is now intense: deep learning specialists with little or no industry experience can expect considerable financial remuneration for their knowledge [7]. The advent of automated deep learning platforms will significantly address this, although awareness of best practice for machine learning, as well as medical statistics, will still be essential.

Challenges of Deep Learning for Clinical Researchers

Prior to the introduction of automated deep learning platforms, a number of barriers existed to the development of deep learning systems in medicine.

Highly Specialized Technical Expertise

Until recently, sophisticated technical expertise was a fundamental requirement for deep learning research in healthcare. Following the triumph of

Compute Resources

Access to significant computing resources, in particular graphical processing units (GPUs), is also a prerequisite for state-of-the-art deep learning. However, most large clinical datasets reside in hospitals where, aside from a few major centers, there is little if any access to GPUs. Since 2012, the amount of computing resources used for major advances in AI has grown by more than 300,000 times. This equates to doubling of requirements every 3.5 months – by comparison Moore’s Law (relating to the number of transistors on an integrated circuit) has an 18-month doubling period.

It is estimated that each of these major advances employs hardware that costs in the single digit millions of US dollars to purchase. AlphaGo Zero and AlphaZero from DeepMind are the most visible public examples of this trend, but many other applications at this scale are now algorithmically possible. More recently, OpenAI has released a series of language models called Generative Pre-trained Transformers (GPTs). These models have drawn criticism for a number of reasons [8, 9], one of which is the sheer size of the models. The authors report that GPT-3 has a capacity of 175 billion parameters [10]. As progress with GPUs reaches a ceiling, there is also likely to be a move toward AI-specific integrated circuits, such as tensor-processing units (TPUs) (designed by Google) and intelligence processing units (IPUs) (designed by the UK company, Graphcore). Thus, it is already clear that it will be difficult for small research groups, working alone in hospital and university settings, to accommodate these huge financial costs and rapidly evolving landscape. Automated deep learning platforms are typically hosted on major public cloud infrastructure. These systems offer the potential for flexible, scalable access to compute resources, as well as providing the potential for unlimited storage. As these systems become more widespread, the cost of credits on these systems is also likely to reduce.

Large, Well-Curated Datasets

The third major challenge for AI-enabled healthcare is the requirement of deep learning systems for large, well-curated datasets. The advent of automated deep learning addresses this, at least in part, by facilitating training of models within a hospital setting, where the best clinical datasets reside. However, the requirement for meticulous labelling of such dataset for supervised learning remains a challenge. The various automated deep learning platforms have begun to address this. The recent release of services, such as Google Cloud AutoML Vision Human Labeling, Clarifai Scribe Label, and Amazon Automate Data Labeling, has offered a promising solution to

these issues. AutoML Vision Human Labeling will label datasets once provided with comprehensive labelling guidelines [11]. Amazon Automate Data Labeling [12] uses a combination of human workers and active learning to automatically label datasets. These services are still in their early stages, and so further work will be required to determine their clinical applicability. The authors have not yet had the opportunity to explore the usability of these features but acknowledge the potential they offer to further increase accessibility of deep learning to smaller research groups without the manpower required to label large datasets.

Data Protection and Privacy

Other barriers to data use include strict data governance and hospital ethics committees which must give appropriate approvals. For many projects, the data must also be de-identified through either anonymization or pseudonymization. Anonymous information is defined as information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is no longer identifiable [13, 14]. Pseudonymization is defined within the GDPR as the processing of personal data such that it can no longer be attributed to a specific data subject without the use of additional information, such as a key [14]. Yet with “de-identified” or pseudonymized data, there remains the risk of re-identification, and this must be taken into account with as much risk mitigated as possible [15]. Given these challenges, the number of publicly available datasets has played a major role in health data research. However, the widespread use of a small number of datasets may result in models which do not generalize well to specific populations, diseases, or outcomes [16]. Fortunately, there are a number of cloud-based tools that address the challenge of de-identification of data such as Google Cloud Healthcare application programming interface (API), Amazon Comprehend Medical API, and Microsoft Sensitive Data Discovery and De-Id Tool. An API, essentially,

allows multiple applications to interact with each other. It is likely that these features will play a significant role in the future of automated deep learning clinical research, thus reducing the dependency on publicly available datasets.

Principles

Automated machine learning describes methods for dataset management, algorithm selection, and optimization of model hyperparameters. Quanming et al. describe these methods as machine learning *tools*, accurately conveying the residual requirement for computer expertise [5]. Examples of such applications within the machine learning industry include Auto-sklearn, H2O AutoML, Auto-keras, and Auto-Weka [17]. In this chapter, we focus more specifically on automated deep learning platforms, a subset of this field where limited if any coding experience is required (i.e., the platforms typically offer drag-and-drop graphical user interfaces for model development). These platforms are typically based on a recent advance in machine learning, termed neural architecture search.

Neural architecture search aims to make the model design process easier and more accessible. Described by Google CEO, Sundar Pichai, as “neural nets to design neural nets,” [18] neural architecture has received a lot of scientific and media attention. Neural architecture search is typically achieved using one of two methods: [1] reinforcement learning algorithms and [2] evolutionary algorithms. The former forms the basis of the commercially available automated deep learning platforms. Zoph and Le described a reinforcement learning approach to neural architecture search [19]. This seminal work is based on the observation that the structure and connectivity of an artificial neural network can be typically specified by a variable-length string (i.e., a linear sequence of characters). It is therefore possible to use a “controller” neural network to generate such a string and then to train the neural network specified by the string – the “child” network – on the real data (Fig. 2). The resulting network will ultimately provide an accuracy measure obtained on a validation set. This accuracy signal can then be used to update the controller network. As a result, the successive iterations of the controller network will give higher probabilities to model architectures that

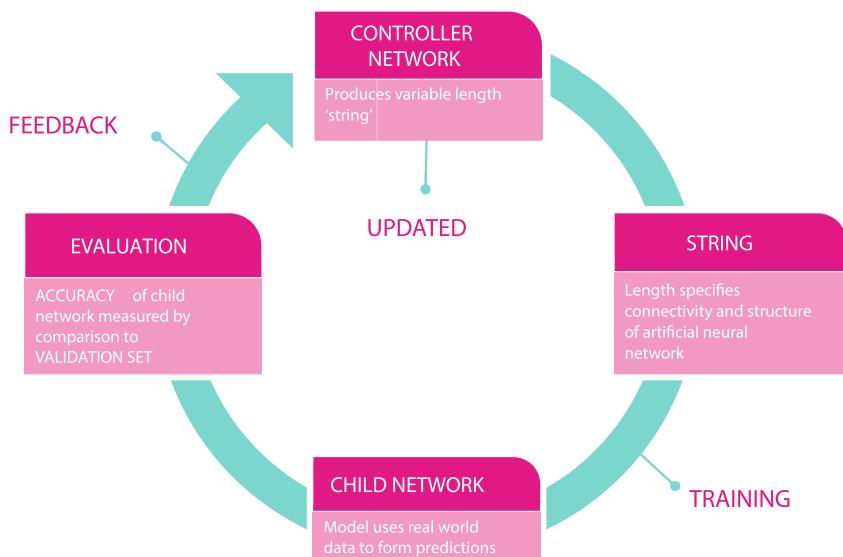


Fig. 2 Illustration of the “reinforcement learning algorithms”

receive high accuracies – i.e., it will learn to improve its search over time. In their initial 2017 work, Zoph and Le apply this approach to two heavily benchmarked deep learning datasets: [1] the CIFAR-10 image recognition dataset and [2] the Penn Treebank language modelling dataset. In both cases, they show that their approach can generate models that achieve accuracies on a par with the state-of-the-art handcrafted models. In April 2018, they extended this work to achieve similar performance on large, more challenging datasets (e.g., ImageNet) – a novel approach that they term NASNet [20].

Although neural architecture search has undeniably resulted in significant advances, it is still largely dependent on expert design and restrictive search spaces. Evolutionary algorithms offer an alternative approach to automating the discovery of deep learning algorithms. Based on the process of biological evolution, these algorithms use mathematics as the building blocks [21]. The approach typically starts with an initial population of one thousand simple models and continues through a number of steps. Each model is a trained network. Initially, a pair of networks are randomly selected to carry out a specific machine learning task. The network that performs more accurately is selected as a *parent*, while the other network is *killed*, analogous to natural selection. This *parent* network is mutated to create a *child* network which is trained, evaluated, and then submitted back into the population. The mutation applied to the *parent* network is selected at random from a prespecified set of actions chosen to be similar to those a human expert may use when designing a network [22, 23]. With the continuation of this process, the population evolves. The practice of selecting the *fitter* of the two models is described as “tournament selection” [24] or non-aging evolution. Alternatively, Real et al. propose the process of aging evolution whereby at each step, the older of the two networks is killed. This avoids the potential pitfall of models reaching high accuracy by chance and instead forces the architecture to re-train each time it is inherited [22].

Initial Work Utilizing Automated Deep Learning in Medicine

In 2019, we reported our initial experiences using automated deep learning for medical imaging applications [25]. Using five publicly available, open-source datasets, members of our research group with no coding expertise were able to develop medical image classification models with comparable performance to state-of-the-art bespoke deep learning models. We further expanded on this work to closely review and compare a number of the commercially available cloud-based platforms using four medical imaging datasets to create image classification algorithms. In this study, we found a difference in performance when comparing both the different platforms and comparing the various models against different datasets, with the best outcomes detected using the OCT image modality. It was also noted that certain platforms were unable to process large imaging datasets [26]. Kim et al. explored the use of AutoML Vision (Google) to classify pachychoroid disease using a dataset of 783 ultrawide-field (UWF) indocyanine green angiography (ICGA) images [27]. They constructed two models using the platform. Using the original dataset, Model 1 was trained to distinguish between pachychoroid eyes and non-pachychoroid eyes. The UWF ICGA images of the left eyes were then horizontally flipped to obtain image orientation of the right eye. Model 2 was trained with the flipped images in order to increase the diagnostic performance. On evaluation, it was reported that Model 2 outperformed physicians in terms of precision, accuracy, and kappa score. Of note, the authors comment that the “platform’s self-evaluation yielded better accuracy and recall when compared to the performance of the deployed model applied on the test sets.” This emphasizes the importance of robust validation and external evaluation. Zeng et al. examine the use of AutoML Vision for the identification of invasive ductal carcinoma using a public dataset of 278,124 labeled histopathology images [28]. As the original dataset was heavily imbalanced, augmentation was performed by “rotating a large portion of the positive image

samples” resulting in a total number of 378,215 labeled images. The AutoML Vision model demonstrated an average accuracy of 91.6% as measured by area under precision-recall curve. The authors also reported that their model showed superior performance compared to the results of bespoke models published in earlier studies. In the near future, it may be possible for a clinician to develop a highly effective model that can be implemented into clinical practice in a manner that is applicable to the specific population without reliance on an artificial intelligence expert.

Automated Deep Learning Process

Although specific features vary across platforms, there is a common process behind the creation of an automated deep learning model. Most platforms (Amazon, Apple, Clarifai, Google, and Microsoft) allow for creation of image classification, segmentation, and object detection models. Some platforms support classification models with both multi- and single-label approaches. For each platform, the graphical user interface (GUI) contains three main components – data upload, data visualization, and model evaluation. We will now describe the process behind each part in detail as it relates to image classification models (Fig. 3).

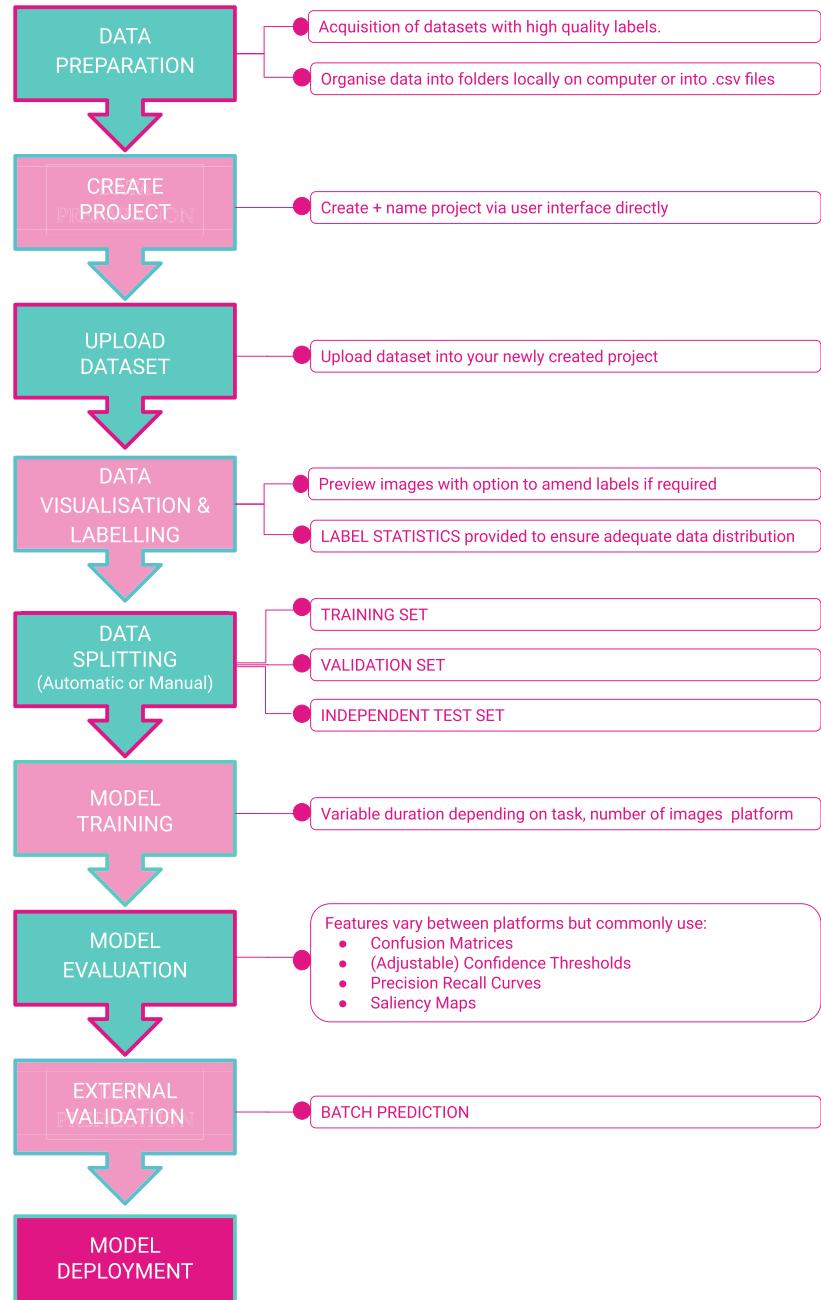
Data preparation, similar to bespoke deep learning projects, must be carried out initially through the acquisition of datasets with high quality labels, navigation of data governance, and organization of the dataset into either folders locally on the computer or with spreadsheet .csv files. The project can be created and named via the user interface directly. The dataset can then be uploaded to the new project. The majority of platforms require for the dataset to be pre-arranged into labelled folders with labels assigned accordingly once uploaded (Amazon, Clarifai, MedicMind, Microsoft). Google allows for data to be converted locally into .csv files and uploaded onto a cloud storage bucket. The images can then be annotated either via the .csv files containing labels or through the user interface. MedicMind allows for data to be manipulated

locally via .csv files, although it does not offer management through a cloud bucket. The Amazon platform can also be linked to a bucket with images labelled accordingly. As Apple does not use cloud computing, labels are assigned based on local folders divided by label.

Once the dataset is uploaded, the images can be previewed with the option of amending any labels. Label statistics are provided to ensure adequate distribution. This is an important process as imbalanced data can be problematic for model training. Google requires a minimum of ten images per label. Amazon Rekognition allows bounding boxes to be assigned, with multiple labels per image. The datasets are divided into three groups – a training set, a tuning set, and an independent test set. Usually 60–80% of the images are allocated to a training set and thus used to set the neural network model parameters. Ten to twenty percent of the images will be allocated to a validation set which will be used to optimize the parameters of the models. Finally, 10–20% of cases will be set aside in independent test sets – this will be used to assess how well the best model from the validation set will generalize to real-world data. Of note, images should be separated at the patient level, to ensure that data acquired from the same patient will not be present in both training and test sets. The training data should be as close as possible to the data on which predictions are to be made. At times, the optimal model parameters used in the validation stage can result in biased metrics. The test dataset provides the training process with an unbiased assessment of the quality of the model. Common issues that arise can include imbalanced data, bad splits, and data leakage [3, 29, 30]. The leakage of data from training to validation sets can result in falsely increased performance. This is something clinicians and researchers must be cognizant of when creating models. The datasets can be split automatically, or alternatively, Amazon, Apple, Google, and MedicMind allow for custom division. Furthermore, Apple uniquely offers the option to perform data augmentation, including cropping, flipping, rotation, and the addition of noise and blur.

Following completion of data preparation and upload, the model is ready to be trained.

Fig. 3 Illustration of the automated deep learning process



The goal of the training process is to enhance the model parameters, so that when a test image is input, the error rate remains at a minimum [31]. The majority of platforms do not have computer system requirements as they can train and evaluate the model using cloud hosted GPUs. Apple runs locally limiting the user to

their available computational resources. Google and Microsoft allow the number of hours for training to be selected, although the model will automatically stop training once it stops improving. Once the model has completed training, detailed statistics for the model will be available.

The results can be reviewed in the evaluation page. Evaluation features vary between platforms. Confusion matrices are generated by Apple, Clarifai, Google, and MedicMind, demonstrating true positives and false negatives. Threshold is provided by Amazon, although it cannot be adjusted. Clarifai, Google, and Microsoft allow the confidence threshold for prediction to be moved to generate new precision and recalls for that confidence. This is useful when evaluating varying operating points for different use cases. For example, if a clinician is to develop a screening test requiring high sensitivity, the confidence threshold will be low in order to create a high sensitivity but result in lower positive predictive value and specificity. Precision recall curves are provided by Amazon, Clarifai, and Google. Clarifai additionally provides ROC curves for each label. Saliency maps are produced by MedicMind which are designed to highlight the components of the input images that are important to the algorithm prediction. They are usually presented as a heatmap with the hotter areas corresponding to the part of the image contributing the greatest impact to the final prediction [32]. MedicMind also allows for a .csv file to be downloaded of the final classification performance. Only Google and Microsoft allow for the final model to be downloaded. Finally, external validation is facilitated only by Google and MedicMind platforms. Batch prediction allows for large amounts of data to be used to ultimately assess the model. Google also allows for images to be uploaded directly onto the graphical user interface for prediction, but this is limited to ten images, making batch prediction a crucial process [26].

Limitations of Automated Deep Learning

Despite the promise of automated deep learning, many of the limitations of bespoke deep learning models also apply. In addition, this is an emerging area so there are a number of platform-specific limitations also.

General Limitations

Explainability

“The strength of machine learning, but also one of its vulnerabilities, is the ability of models to discern patterns in historical data that humans cannot find” [3]. This is often referred to as the “black box” in artificial intelligence. Even if the underlying mathematical principles of a model is understood, it is difficult and often impossible to determine the inner workings to ascertain how and why it made a certain decision. This lack of explainability is intensified by automated deep learning platforms, where the underlying model architecture is opaque.

At present, in order for a new pharmaceutical drug or medical device to be approved for patient use, it must undergo rigorous testing. This must also be the case when it comes to implementing automated deep learning models to ensure the safe development, use, and monitoring of systems. The term “AI chasm” has been used to describe the fact that accuracy does not necessarily represent clinical efficacy [33]. Healthcare is one of the predominant fields in which explainability is essential, given the potential for devastating consequences from prediction errors. The ability to interpret the algorithm could also enlighten researchers to patterns not previously detectable by humans and lead the way toward new scientific advances [30].

Generalizability and Bias

Although automated deep learning has somewhat alleviated the requirement of advanced computer expertise, it cannot provide a “one-size-fits-all” model. Clinicians must be aware of the dangers of overfitting. Between training sites, there can be differences in equipment, coding definitions and, electronic health record systems. There are also minor differences between the various commercially available automated deep learning platforms in terms of training and evaluation. Furthermore, the issue of bias must be addressed. In some instances, the dataset on which the model is trained can reflect societal inequalities, including bias toward those who contributed the most data [29]. Model bias occurs when the model is

selected to best represent the majority, potentially resulting in underperformance on underrepresented groups. Model variance occurs due to inadequate data from minorities. Discriminatory bias was demonstrated by Esteva et al. where the deep learning algorithm developed for the detection of skin cancer could classify skin lesions with similar accuracy to that of board-certified dermatologists in fair-skinned patients but showed underperformance on images in skin of color. This is largely based on the fact that the model was trained on open datasets of predominantly fair-skinned patients emphasizing the critical need for a diverse dataset that accurately represents the population it is intended for [34]. It is possible that the increased accessibility that comes with automated deep learning platforms may further exacerbate this problem. Clinicians must be acutely aware of this issue when developing algorithms and question whether they are appropriate for their specific population.

Automated Deep Learning-Specific Limitations

In our previous work, assessing six commercially available automated deep learning platforms, we noted a difference in performance between our automated deep learning models and bespoke models published in the literature using the same datasets. We postulate the reasons behind this to be “automated deep learning’s lack of task-specific image augmentation pre-processing, inability to specify task-specific base models for transfer learning approaches, and the inherent performance variations resulting from bespoke model construction and tuning” [26]. Even for those with some technical background, the automated deep learning platforms have limitations. There is no flexibility in choosing between models to train – AutoML Vision uses a mobile neural architecture search with no other available option. Furthermore, the user is frequently not provided with information on which model architecture is being utilized. This is an obvious barrier to

clinical implementation. As discussed, similar basic tasks are implemented on all platforms, e.g., image classification, object detection, and segmentation. However, other potentially useful model designs are frequently not available. For example, regression models are not supported by Google AutoML Vision or Amazon Rekognition for predicting continuous variables. Semantic segmentation was previously only possible with Amazon Rekognition but has recently become available with Google’s Unified AI Platform. Additionally, it is not possible to specify task-specific base models for transfer learning approaches on automated deep learning platforms at present. Once the model has been trained, there are differences in the evaluation and result metrics provided by various platforms. Saliency maps are provided by MedicMind; however, these maps should be interpreted cautiously as it has been shown in some studies that they can lack sensitivity to the model and the insights obtained from them may be subjective [32, 35]. Confusion matrices are frequently not provided, which are needed to calculate other clinically relevant metrics such as specificity.

External validation is an essential step when training an automated deep learning model and one that must be carried out prior to implementation. The ability to perform batch predictions for the purpose of external validation is only available with Google and MedicMind platforms. Models that are unable to be externally validated cannot be implemented without assessing their performance on local populations. It is not possible to export Google AutoML Vision cloud models at present, although it is possible to export edge models for mobile devices. Even with comprehensive external validation, implementation will require an ongoing performance monitoring process. As automated deep learning becomes accessible to a broader range of clinicians and researchers, it is essential that the principles discussed above are adhered to in order to maintain responsible and safe AI that is of benefit to the population it is being used on.

Future Directions of Automated Deep Learning

Use Cases of Automated Deep Learning

Undoubtedly, there are significant barriers to overcome before an automated deep learning model can be implemented into the direct clinical care pathway. Nonetheless, it is possible for such models to considerably benefit patients indirectly, in ways that do not require extensive regulatory approval. This is particularly relevant in the field of clinical research. In order to create algorithms in healthcare, close collaboration between clinicians and data scientists is required. Automated deep learning looks set to drastically alter the clinical research landscape affording clinicians the independence to create their own models. Automated deep learning models developed as part of a research toolkit could liberate clinicians from the time-consuming tasks associated with research. For example, a model could be trained to distinguish between the left and right eyes from a folder of unlabeled retinal images rapidly leading to a cleaned and usable dataset. Moreover, a model could be trained on a smaller dataset to then label large unlabeled datasets, saving the clinician considerable time and effort. Automated deep learning models could aid the clinical researcher by allowing for a model specifically representative of the target population to be designed. Alternatively, models could be created within an institution to assist with image quality assurance and improvement. An automated deep learning model created to monitor the quality of retinal photographs could be deployed to run continuously and detect any decreases in image quality. This could help alert hospital management to specific problems, allowing for rapid identification and resolution. Additionally, an automated deep learning model that could automatically analyze certain imaging modalities to search for specific features would greatly enhance identification of patients suitable for clinical trials. This would

help clinical researchers to rapidly identify a patient cohort and reduce the time taken to commencing the trial.

Packaging and Deployment of These Models

The efficient packaging and deployment of these models is a crucial step in the pathway toward real-world implementation of automated deep learning. Some platforms allow models to be served through the cloud via APIs. Despite the previously mentioned advantages of cloud computing, remaining issues include latency, scalability, and privacy. These issues particularly become problematic when applied to clinical uses, which depend on quick transfer of data, robust privacy features, and high bandwidth. A possible option is to deploy the automated deep learning algorithm via a local edge model [36]. By keeping data local, the issues of decreased bandwidth, greater latency, privacy, and reliability concerns are potentially avoided [37]. Moreover, edge models do not require continuous Internet for the model to run, making them particularly advantageous in under-resourced areas. Bellemo et al. describe how artificial intelligence can directly enhance a diabetic retinopathy screening program in Zambia, where there is a limited number of ophthalmologists for a large population. Nonetheless, they report the affordability of the model as a limitation, particularly with poor Internet infrastructure [38]. A similarly effective model, created using automated deep learning and deployed on an edge model, could greatly reduce costs and allow for scarce resources to be redistributed. The recent COVID-19 pandemic has further emphasized the crucial role automated deep learning algorithms could play using edge computing. Deployment of real-time and contactless screening tools could aid with rapid diagnosis and the detection of clinical deterioration, allowing for limited resources to be directed appropriately [39]. With the approaching widespread availability of 5G technology, greater bandwidth and lower latency are predicted. The

new 5G base stations are expected to “benefit network science and efficient management by providing cost-effective and efficient systems to support massive data collection” [40]. Additionally, the computational power of smartphones is increasing exponentially – a phone with 512GB of RAM today has seven million times more than that of the Apollo 11 guidance computer [41]. All of these reasons emphasize the role edge-based computing is likely to play in the future of automated deep learning.

Conclusion

Automated deep learning may serve to democratize AI in healthcare. At present, machine learning is challenging due to limited resources. This is particularly relevant to clinicians and clinical researchers. Automated deep learning allows those without coding experience to create their own models. The true potential of this emerging field is yet to be seen as researchers take advantage of the ever-expanding amount of clinical data and begin to create their own models for their own use cases to enable fundamental improvements in patients’ lives.

References

- Metz C. Building AI that can build AI. The New York Times Çevrimiçi (Erişim, 4 Şubat 2018). 2017. <https://www.nytimes.com/2017/11/05/technology/machine-learning-artificial-intelligence-ai.html>
- LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.
- Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347–58.
- Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56.
- Yao Q, Wang M, Chen Y, Dai W, Li Y-F, Tu W-W, Yang Q, Yu Y. Taking human out of learning applications: a survey on automated machine learning. arXiv:181013306v4 [Internet]. 2019 Dec 16. <https://arxiv.org/abs/1810.13306>
- Economist T. Million-dollar babies. The Economist, Apr 2nd. 2016;9.
- Metz C. AI researchers are making more than \$1 million, even at a nonprofit. NY Times. 2018.
- Toews R. Deep learning’s carbon emissions problem [Internet]. 2020. <https://www.forbes.com/sites/robtoews/2020/06/17/deep-learning-climate-change-problem/?sh=500edbee6b43>
- Marcus G, Davis E. GPT-3, bloviator: OpenAI’s language generator has no idea what it’s talking about. Technol Rev. 2020. <https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/>
- Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners [Internet]. arXiv [cs.CL]. 2020. <http://arxiv.org/abs/2005.14165>
- Human labeling [Internet]. AutoML Vision Guides. 2020 [cited 2021 Jan 13]. <https://cloud.google.com/vision/automl/docs/human-labeling>
- Automate Data Labeling [Internet]. Amazon Web Services. [cited 2021 Jan 13]. <https://docs.aws.amazon.com/sagemaker/latest/dg/sms-automated-labeling.html>
- Act A. Health insurance portability and accountability act of 1996. Public Law. 1996;104:191.
- European Union. European data protection law: general data protection regulation 2016. CreateSpace Independent Publishing Platform; 2016. 130 p.
- Na L, Yang C, Lo C-C, Zhao F, Fukuoka Y, Aswani A. Feasibility of reidentifying individuals in large national physical activity data sets from which protected health information has been removed with use of machine learning. JAMA Netw Open. 2018;1(8):e186040.
- Khan SM, Liu X, Nath S, Korot E, Faes L, Wagner SK, et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. Lancet Digital Health. 2021;3(1):e51–66.
- Truong A, Walters A, Goodsitt J, Hines K, Bruss CB, Farivar R. Towards automated machine learning: evaluation and comparison of AutoML approaches and tools. In: 2019 IEEE 31st international conference on tools with artificial intelligence (ICTAI). 2019. p. 1471–9.
- Pichai S. AI first [Internet]. Google Input/Output; 2017 May 17; California. <https://www.youtube.com/watch?v=CNLVZjBE08g>
- Zoph B, Le QV. Neural architecture search with reinforcement learning [Internet]. arXiv [cs.LG]. 2016. <http://arxiv.org/abs/1611.01578>
- Zoph B, Vasudevan V, Shlens J, Le QV. Learning transferable architectures for scalable image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 8697–710.
- Real E, Liang C, So D, Le Q. AutoML-zero: evolving machine learning algorithms from scratch. In: International conference on machine learning. PMLR; 2020. p. 8007–19.
- Real E, Aggarwal A, Huang Y, Le QV. Regularized evolution for image classifier architecture search. AAAI. 2019;33(01):4780–9.

23. Real E, Moore S, Selle A, Saxena S, Suematsu YL, Tan J, et al. Large-scale evolution of image classifiers. In: Precup D, Teh YW, editors. Proceedings of the 34th international conference on machine learning. Proceedings of machine learning research, vol. 70. International Convention Centre, Sydney: PMLR; 2017. p. 2902–11.
24. Goldberg DE, Deb K. A comparative analysis of selection schemes used in genetic algorithms. In: Rawlins GJE, editor. Foundations of genetic algorithms. Elsevier; 1991. p. 69–93.
25. Faes L, Wagner SK, Fu DJ, Liu X, Korot E, Ledsam JR, et al. Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study. *Lancet Digit Health.* 2019;1(5):e232–42.
26. Korot E, Guan Z, Ferraz D, Wagner SK, Zhang G, Liu X, et al. Code-free deep learning for multi-modality medical image classification. *Nat Mach Intell.* 2021;3:288.
27. Kim IK, Lee K, Park JH, Baek J, Lee WK. Classification of pachychoroid disease on ultrawide-field indocyanine green angiography using auto-machine learning platform. *Br J Ophthalmol [Internet].* 2020. <https://doi.org/10.1136/bjophthalmol-2020-316108>.
28. Zeng Y, Zhang J. A machine learning model for detecting invasive ductal carcinoma with Google Cloud AutoML Vision. *Comput Biol Med.* 2020;122: 103861.
29. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med.* 2019;25(9):1337–40.
30. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 2019;17(1):195.
31. Zaharchuk G, Gong E, Wintermark M, Rubin D, Langlotz CP. Deep learning in neuroradiology. *AJNR Am J Neuroradiol.* 2018;39(10):1776–84.
32. Arun NT, Gaw N, Singh P, Chang K, Hoebel KV, Patel J, et al. Assessing the validity of saliency maps for abnormality localization in medical imaging [Internet]. arXiv [cs.CV]. 2020. <http://arxiv.org/abs/2006.00063>
33. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. *NPJ Digit Med.* 2018;1:40.
34. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 2017;542(7639):115–8.
35. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps [Internet]. arXiv [cs.CV]. 2018. <http://arxiv.org/abs/1810.03292>
36. Chen J, Ran X. Deep learning with edge computing: a review. *Proc IEEE Inst Electr Electron Eng.* 2019;107(8):1655–74.
37. Merenda M, Porcaro C, Iero D. Edge machine learning for AI-enabled IoT devices: a review. *Sensors [Internet].* 2020;20(9). <https://doi.org/10.3390/s20092533>.
38. Bellemo V, Lim ZW, Lim G, Nguyen QD, Xie Y, Yip MYT, et al. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. *Lancet Digit Health.* 2019;1(1):e35–44.
39. Greco L, Percannella G, Ritrovato P, Tortorella F, Vento M. Trends in IoT based solutions for health care: moving AI to the edge. *Pattern Recogn Lett.* 2020;135:346–53.
40. McClellan M, Cervelló-Pastor C, Sallent S. Deep learning at the mobile edge: opportunities for 5G networks. *NATO Adv Sci Inst Ser E Appl Sci.* 2020;10(14):4735.
41. Kendall G. Apollo 11 anniversary: could an iPhone fly me to the moon. *Independent.* 2019.



AI in Musculoskeletal Radiology

34

Stefan Nehrer, Philip Meier, Matthew D. DiFranco,
Zsolt Bertalan, and Richard Ljuhar

Contents

Introduction	488
Machine/Deep Learning	488
Musculoskeletal (MSK) Diseases: A Rapidly Rising Global Burden	490
Bringing Disease Management into the Digital Age	491
Musculoskeletal Radiology	491
AI in Computed Radiographs	492
Building and Validating AI Models for MSK Radiographs	493
Quality Control	494
Anomaly Detection	495
IB Lab KOALA: Assessment of the Gonarthrosis Stage	495
IB Lab PANDA: Determination of Pediatric Bone Age	495
IB Lab HIPPO: Measurement of the Hip and Pelvis	496
IB Lab LAMA: Measurement of the Whole Leg	497
AI Disrupts Status Quo of Radiology Reporting	498
Take-Home Message: Artificial Intelligence (AI) in Musculoskeletal Radiology ...	499
References	499

Abstract

S. Nehrer
Donau-Universität Krems, Zentrum für Regenerative
Medizin, Krems, Austria
e-mail: stefan.neher@donau-uni.ac.at

P. Meier · M. D. DiFranco (✉) · Z. Bertalan · R. Ljuhar
ImageBiopsy Lab, Research & AI Development
Abteilung, Wien, Austria
e-mail: p.meier@imagebiopsy.com;
m.difranco@imagebiopsy.com;
z.bertalan@imagebiopsy.com; r.ljuhar@imagebiopsy.com

Digitalization and artificial intelligence (AI) have reached orthopedics and traumatology. AI improves diagnostic accuracy in knee osteoarthritis according to the latest clinical guidelines and facilitates the (radiological) monitoring of the progression of various bone and joint diseases. AI makes it possible to detect early radiological signs and allows conclusions with respect

to progression of the disease and prognosis. By automating the report generation, patient throughput can be increased and the radiologist's workload decreased. AI also reduces the impact of inter-rater variability in the assessment of radiographic X-ray morphologies and creates more standardized outcome measurements.

Keywords

Artificial intelligence · Imaging digital data · Machine learning · Deep learning · Neural networks · Muscular skeletal disease

Introduction

Digitalization is a phenomenon that has accompanied us for decades and has spread to many processes in our social life, society, and economy and thus also in medicine. The process of digitalization was already so far advanced that only a few years ago action had to be taken to further develop and finalize the whole concept. With the COVID pandemic, digital processes have anyway permeated our world and largely taken over. This is also the case in the diagnosis of X-ray images in orthopedics/traumatology, where rapid developments of digital image processing and analysis are on the way. However, the diagnosis of X-ray images is still largely carried out manually with narrative image descriptions, which are very subjectively impacted. This is also reflected in the high inter- and intraindividual variability of the findings, and thus, especially in the assessment of diseases such as osteoarthritis, a low accuracy and comparability is present. Despite the fact that we hold five megabytes of image data in our hands in digital X-ray procedures with DICOM formats, these are subjectively assessed on the screen in a quantifying and qualifying manner – actually in the same way as X-ray images were held up to the light and analyzed a hundred years ago. Artificial intelligence (AI) are computer programs – so-called algorithms – extracting and recognizing characteristics and patterns of typical changes caused by

diseases from the digital data material and subsequently learning to make diagnosis on the basis of objective data analysis. This process is called “machine learning” whereby data material and information about the criteria must be uploaded. Deep learning includes the use of deep neuronal networks. In those algorithms, the data learns itself their own characteristics and analyzes and reinforces them in forward and backward loops, so the data itself learns which patterns it contains. This significantly surpasses our limited ability to assimilate, since all digital data enters here and not just limited criteria that we include in the manual diagnosis. The result becomes more objective, more accurate, and almost 100% reproducible. Furthermore, structural analyses can be performed, which cannot be detected with the naked eye – not even with a magnifying glass. As an example, we can see constellations in the sky, connecting particularly visible stars to form a symbolic image. When we look through a high-power telescope, so many stars become visible that we cannot longer identify those constellations, although they are of course still there, but we actually have too much information to coin out the symbolic figure. Therefore, we need digital methods that help us to understand this flood of information and also to interpret this data in order to get a much more objective result, since much more data contribute to the result. AI uses this digital data, learns itself what characteristics and patterns it contains, and then gives this information back to us.

Machine/Deep Learning

Machine learning is a field of computer science and a component of artificial intelligence. Computer programs, which are based on machine learning, can find solutions for new and/or unknown problems by means of algorithms independently. Different types of algorithms are distinguished:

- Supervised learning: Here algorithms are defined by specific examples. Thereby it is

tried to find the solution for further similar problems by the generalization of a solution.

- Unsupervised learning: Here algorithms are processed with arbitrary examples. The goal is to recognize a structure within the dataset.
- Transduction: This method tries to find new solutions based on specific cases.
- Learning to learn: In this method, algorithms draw derivations from experience already gained.
- Developmental learning: Here, a software learns almost independently through exchange with human “teachers” [1].

Deep learning is similar to machine learning but goes beyond it. It uses neural networks. The goal of deep learning is to process and analyze large amounts of data. By using neural networks, already existing information can be interpreted and further processed. This allows learned information to be merged with new content and used for future tasks [1] (Fig. 1).

In order for neural networks to recognize a disease on an X-ray image, all characteristics must be presented beforehand, if possible, so that the network can later evaluate new X-ray images correctly in practice. Therefore, all different disease stages, patient morphologies, and image qualities

must ideally be present in the training data. Artificially, additional variance can be generated by “augmentation.” To train, a reward system is used that employs numerical optimization. The more the findings are available in numbers and categories, the easier this optimization can be done. The more consistent the training data, the more accurate the training on less data.

Example 1 Two radiologists have different opinions on the same dataset of an image. If we train on this data, the network will reflect the difference in opinion and will perform poorly. A solution would be to have a third radiologist arbitrate and harmonize the disagreements.

Example 2 A radiologist at Stanford started 10 years ago to annotate every X-ray with 1 or 0 if there is an anomaly or not. From the 40,000 consistently found images, extremely accurate anomaly detection networks could now be trained. Conclusion of this case: The more variance there is in the radiological appearance or annotations, the more data is needed.

In recent years, a portfolio of AI-driven decision support tools for musculoskeletal (MSK) disorders has been developed to help radiologists and

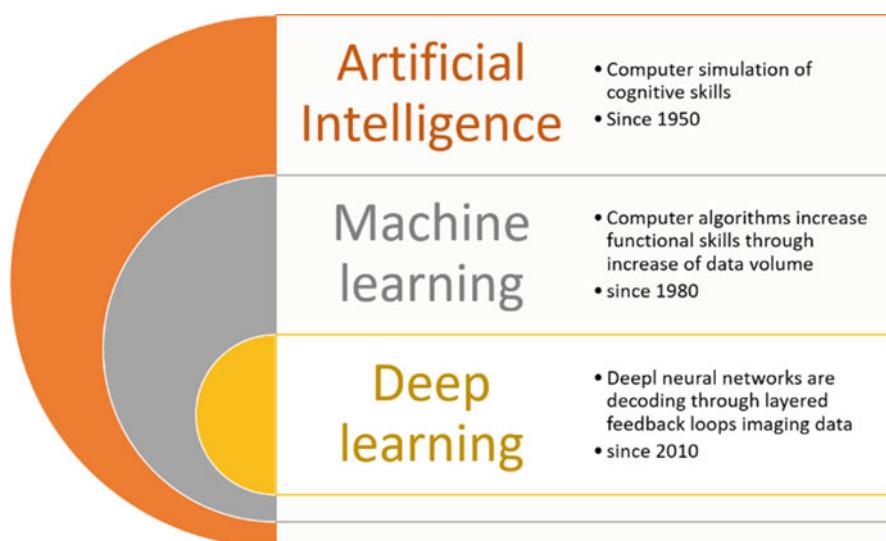


Fig. 1 Deep learning is a special form of machine learning and part of artificial intelligence (AI)

orthopedists quickly and accurately assess skeletal radiographs. Some of these products are already CE certified, such as modules for bone age, knee, whole-leg, and hip assessment, covering the majority of radiologists' and orthopedists' skeletal X-ray workload. Calculations take less than a minute. Usually, the processing is started automatically, and the results are then immediately attached to the original study as a separate series. In the following, the impact of AI in imaging is described and some examples of the application are explained.

Musculoskeletal (MSK) Diseases: A Rapidly Rising Global Burden

The global aging population is currently at the peak in human history. Understanding the healthy, aging population and reducing the socioeconomic impact of age-related diseases are a key research priority for industrialized and developing countries, along with a better mechanistic understanding of the physiology of aging. MSK conditions were the leading cause of disability in 4 of the 6 WHO regions in 2017, affecting approx. 1.3 billion people globally and more so when taking into account additional MSK burden from long-term sequelae of fractures and dislocations.

Musculoskeletal conditions can be classified by more than 150 diagnoses that include, but are not limited to, a single region of muscles, joints, bones, and surrounding tissues such as tendons and ligaments. This includes sudden and short-onset conditions, such as fractures strains and sprains, all the way to slowly progressing lifelong problems with chronic pain and disabilities, all of which significantly reduce the quality of life.

One out of five people (20%) in the EU-28 suffered from a chronic back or neck disorder in 2019.

The prevalence and impact of MSK conditions is predicted to rise as the global population ages.

Needless to say that this scenario is not accepted by the scientific community and a number of recent advances have led to many significant changes in treatment decisions and health pathways. However, even though in some cases

these new approaches have enabled people to regain function and mobility, factual cures for the most common disorders remain elusive.

Nonetheless, the clinical goal of preventing disability is attainable.

Among the most common MSK disorders, osteoarthritis (OA) presents (the most common type of arthritis) itself as affecting 237 million people globally, expanding at a rate of 33% within the last 10 years. It is the fastest-growing cause of disability worldwide. In Europe and the United States alone, more than 70 million people have some kind of physician-diagnosed form of OA.

The US Centers for Disease Control (CDC) presented numbers from 2013 to 2015 that showed an estimated 54.4 million US adults (22.7%) had been physician-diagnosed to have some form of arthritis, rheumatoid arthritis, gout, lupus, or fibromyalgia.

Furthermore, there is an estimated projection of 78 million US adults to suffer from physician-diagnosed arthritis, by 2040.

MSK conditions affect people across the life course in all regions of the world. While the prevalence of MSK disorders increases with age, younger people are also affected, often during their peak income-earning years, accounting for the greatest proportion of lost productivity in the workplace.

In Germany, for example, MSK disorders accounted for €17.2 billion of production loss in 2016 and €30.4 billion in loss of gross value added. This represents 0.5% and 1.0% of Germany's gross domestic product (GDP), respectively. (European Risk Observatory Report—Work-related musculoskeletal disorders: prevalence, costs and demographics in the EU.)

The US CDC estimates that back in 2013, the total national arthritis-attributable (direct and indirect) medical care costs and earnings losses among adults with arthritis were \$303.5 billion or 1% of the 2013 US GDP. OA is also among the most expensive conditions to treat when joint replacement surgery and elaborate post-surgery rehabilitation are required (2015: more than 600,000 end-stage surgeries annually). In fact, OA was the second most costly health condition treated at US hospitals in 2013. In

that year, it accounted for €16.5 billion, or 4.3%, of the combined costs for all hospitalizations. Future projections show that the (indirect and direct) costs for OA are rapidly rising to become a one of the leading burdens of healthcare systems.

In the United States, more than 53 million people either already have osteoporosis (OP) or are at high risk due to low bone mass (<https://www.bones.nih.gov/health-info/bone/osteoporosis/conditions-behaviors/osteoporosis-arthritis#fyi>). Osteoporosis is known as a silent disease because it can progress undetected for many years without symptoms until a fracture occurs. OP is diagnosed by a bone mineral density test using an often challenged radiological method called DXA (dual energy absorption) which provides different classifications based on standardized scores (T-/Z-score, Singh Score, FRAX Score, etc.). The main concern with this method is the low sensitivity in detecting patients at fracture risk. Digital X-rays provide a cost-effective alternative or CT imaging for a high image-resolution assessment of (e.g.) vertebral fractures.

Further MSK-related traumatic incidents include injuries such as cartilage tears or bone fractures due to accidents or excessive stress. For children and adults younger than age 45, trauma accounts for an estimated 79,000 deaths each year, and in the United States, trauma is the single most important cause of potential years of life lost for persons under age 65. Trauma-related injuries account for more than 400 million medical imaging procedures every year, including mostly X-rays and CT (for bone-related examinations) as well as MRI (for cartilage-related examinations).

Bringing Disease Management into the Digital Age

The key for a successful disease management is an accurate identification of relevant imaging biomarkers (bone/cartilage anomalies, biomechanical parameters, clinical scores, etc.) and to monitor the progression using established (and novel) imaging biomarkers in order to adjust the treatment regime.

Currently, the existing paradigm is purely reactive. Disease diagnostics are based on manual assessments, often dependent on personal experience, time allocation, and medical specialty. Result documentations are often based on verbal descriptions rather than standardized disease report findings. Moreover, as a result of the growing number of cases, physicians are faced with an ever-increasing number of low-risk, high-volume MSK image data to analyze.

This results in a combination of missing critical disease parameters (thus discharging a patient too early or not applying the optimum treatment pathway), while at the same time there is not enough time/resources available to provide the required diligence in image reading and result documentation. The consequences are drastic: a rising number of patients suffering from inadequate treatments and healthcare systems overwhelmed by avoidable procedures.

Musculoskeletal Radiology

Musculoskeletal (MSK) radiology is a subspecialty of diagnostic radiology focused on imaging of the bones and joints, as well as the spine and soft tissues. MSK imaging is used to gain a better understanding of human anatomy and to evaluate MSK diseases using diagnostic methods and criteria developed from that understanding.

Within the context of AI, MSK radiology can be broadly classified into 2D and 3D imaging modalities, with each presenting unique challenges and opportunities.

Recent advances in AI have arisen in large part due to breakthroughs in deep learning, a branch of machine learning which uses multilayered models, or networks or models, to transform raw input data into predictions of some higher-level set of features. Deep learning has shown particular promise in the area of computer vision, including image segmentation, object detection and recognition, and scene understanding to name a few.

These computer vision problems are almost exclusively based on 2D color digital photographs taken from the real, physical world. The pixel

representation of digital photos is usually in the RGB space, meaning that there are three color channels (red, green, and blue), and each pixel holds a value between 0 and 255 for each channel, for a total of over 16 million possible colors at any given pixel. This numerical range as well as the spacing of pixels is dictated by the imaging detector used to capture a digital image.

With the success of deep learning on photographic image data, many researchers have adapted models to fit the paradigm of 2D medical imaging. In MSK radiology, this includes computed radiography (CR), a strictly 2D modality, as well as CT and MRI, which are volumetric but nevertheless rely on stacks of 2D images.

CR images are composed of pixels whose values represent the amount of X-ray radiation that hit the detector at a given point for a given exposure time. Unlike digital photographs, CR image pixels hold a single intensity value, although the bit depth of intensities is often higher, with systems producing 12-bit and sometimes 16-bit images. The additional bit depth leads a larger range of values at each pixel and thus higher contrast.

Because CR images are acquired by shooting an X-ray beam at a patient and capturing the signal after it passes through the patient on a flat detector array, the resulting image is a 2D projection of the 3D objects. In addition, since X-rays are attenuated differently depending on the tissue types, with bone being the most “radiopaque,” followed by muscle, low-density muscle, and adipose tissue, CR images are ideally suited for assessment of MSK anatomy, among other uses.

CT and MRI follow a similar principle to CR but represent 3D *in situ* measurements of tissue and produce stacks of images that form a volume. The volumes are composed of voxels, short for “volumetric pixel,” which represent a physical volumetric location within the patient. The intensity of a CT or MRI voxel represents some characteristic of a signal detected at that given position within the patient. Unlike CR imaging, CT and MRI data are not projections, which map directly to a detector array, but rather are reconstructions of signals which pass through the body and are captured by detectors surrounding the patient.

As CT imaging is based on X-ray radiation, the resulting images have similar intensity and contrast characteristics as CR images, with an emphasis on bone and muscle. MRI relies not on radiation but rather on the excitation and relaxation of protons within water molecules in living tissues when in the presence of fluctuating magnetic fields. MRI can be tuned to produce different intensities and contrasts for different tissue types and generally features soft tissue contrast rather than bony structures.

Both CT and MRI are advanced imaging techniques which have the benefits of producing 3D representations of 3D objects, thereby removing the uncertainty in localizing findings and anatomical features in CR projection images. However, CR imaging produces higher-resolution images compared to CT or MRI. A typical CR image has a pixel size of between 0.05 mm and 0.25 mm, whereas CT images have resolutions which typically range from 0.5 mm to 1.0 mm depending on the field of view. For MRI, in-plane voxel resolutions of 1 mm to 2 mm are common. In addition, CT and especially MRI are both susceptible to noise and motion artifacts resulting from the physics involved in each modality as well as the time required to obtain an image.

All of this is to say that, while 3D imaging such as CT and MRI provides clinical advantages owing to the 3D volumetric imaging, their complexity, costs, and requirements result in higher variability of outputs and less overall images acquired in comparison with CR. In terms of AI suitability, the ubiquitousness and ease of CR imaging, along with standardization of imaging protocols and the high contrast features inherent in X-rays, make X-rays the best analog to digital photos when considering applications of deep learning in medical imaging.

AI in Computed Radiographs

As discussed above, CR imaging is well suited to the implementation of AI based on the technical characteristics of X-ray imaging and the capabilities of ANNs. Nevertheless, the motivation for applying AI to MSK radiographs goes beyond

technical compatibility and encompasses the needs of clinicians, patients, and the public. Those needs are best summarized in terms of efficiency, accuracy, standardization, and overall patient well-being. Development of AI-based tools on MSK radiographs should be undertaken with these needs in mind in order to ensure successful adoption by stakeholders.

Implementing AI for MSK applications on CR radiographs can be understood as the automation of specific tasks which are normally performed by radiologists, radiography technologists, and various referring physicians. Those tasks range in scope from image quality control and measurement of anatomical features to detection anomalies like fractures, lesions, and tumors to reporting of more high-level assessments like bone age or osteoarthritis grade. Common to all of these tasks is the requirement for large amounts of annotated example images in order to train and validate AI models designed to automate them to the point of clinical acceptance. The next sections discuss considerations for building and validating AI models, followed by a deeper look at specific problems in MSK radiology where AI automation can have an impact on stakeholder needs.

Building and Validating AI Models for MSK Radiographs

One of the main advantages of focusing on MSK X-rays for building AI models is the widespread, standardized use of CR, which has led to the availability of millions of images for training and testing AI models. Nevertheless, a number of steps should be taken in order to obtain data which is appropriate for the desired task, and to ensure patient privacy and safety is maintained. First and foremost, ethical approval from any and all institutional review boards (IRBs) for the use of retrospective imaging should be sought and obtained prior to carrying out model training or validation. In particular, patient-informed consent must be obtained if necessary, or alternatively, a waiver of the requirement for informed consent should be obtained.

In addition, IRB approval should be contingent on the protection of patient privacy, which will usually entail the pseudonymization or anonymization of imaging data prior to release to the investigator(s). Special attention should be paid to ensuring that protected health information (PHI) is removed from images and that, where required, personal identifiable information (PII) is also removed. Again, any and all relevant regulations should be followed (e.g., HIPPA in the United States, GDPR in Europe).

Pseudonymization refers to the blinding of PHI and/or PII from the investigators, but maintaining the ability to link back to that information. The investigators would receive data that is coded with an identifier which, when given to a designated representative at the data source (e.g., a study nurse or PACS administrator), would allow them to reidentify the patient and associate them with the image data. Pseudonymization allows for investigators to carry out studies with knowledge of clinically relevant information, without being aware of the identity of study subjects.

Anonymization is a means of completely removing the association between imaging data and the patient identity. Anonymization is the safest method of sharing data with an investigator from the standpoint of patient privacy, but it presents limitations in terms of traceability of image data. Training data used to train AI models can be anonymized as long as the requirement for patient traceability is not applicable. For example, training data might be obtained from Hospital A to build a model for predicting the presence of rib fractures from thoracic X-rays. The image would be fully anonymized and sent to the investigators, who would then obtain annotations of the images externally and train their models.

Where traceability becomes a concern is when the model is to be validated on an independent set of data, also known as standalone performance testing, which is generally required for regulatory clearance. If the standalone performance data is also sourced from Hospital A, then the anonymization of the training data may make it impossible to ensure that image sets obtained for standalone testing do not contain images that were part of the model training.

The ultimate goal of building AI models for MSK X-rays is usually to automate a task related to the visual evaluation of an image by a trained expert. A successful AI model must therefore account for the variability of image data in terms of the patient population, imaging parameters, anatomical features, and disease manifestations. The most effective way to achieve robustness in an AI model is to provide training data that encompasses the variability expected in the real world and to provide lots of it. Deep learning thrives on more of everything: data, variety, and annotations.

When training AI models, a number of augmentation techniques can also be used on the data at hand to improve robustness. These techniques include applying various filters and transformations to the images which help simulate image quality issues that are typically found in real-world images.

Advances in synthetic data generation have also shown promise for AI in other application domains and have also been used for generating realistic X-ray images for rare conditions. These developments are particularly interesting for AI in medical imaging applications due to the burdens of acquiring patient data and the challenges in building robust datasets. Ultimately, the clinical acceptance of AI models for any medical imaging automation task will rely largely on standalone performance testing on data not involved in the training of the models. For that reason, the inclusion of synthetic data in model training may prove to be effective, so long as the performance testing targets can be achieved.

Acquisition of diverse and plentiful training data will only get you part of the way to building an AI model with high performance. Almost as important is defining the clinically meaningful measurements, assessments, or other outputs which need to be automated. The automated evaluation of osteoarthritis (OA) in standing PA fixed-flexion knee X-rays offers insights into how expert annotations can help shape the performance of the AI models.

Knee OA is assessed using a number of radiographic indicators, some of which are anatomical measurements, in this case joint space width

(JSW), and some of which are based on consensus grading scales, namely, Kellgren-Lawrence grading (KL). JSW measurement involves identifying the minimum distance between the distal femur and proximal tibia, with the aim of estimating the knee cartilage thickness. Given that cartilage is not visible in CR images and that the image itself is a 2D projection of the knee, the JSW measurement itself is at best an estimation of a physical distance.

In this example, one strategy for building an AI model would be to acquire expert annotations from radiologists and orthopedic surgeons who are tasked with drawing their estimation of JSW in the medial and lateral compartments of both knees (four measurements in total per bilateral X-ray). The model could then be trained using the images as inputs and the JSW measurements as outputs.

However, given that the goal is to accurately measure a distance between two surfaces, another strategy would be to segment the individual bones, as well as the joint space, and then measure the distance in between.

Quality Control

MSK imaging targets specific areas of the skeleton with the aim of making diagnostic evaluations based on anatomical, morphological, or contrast assessment. The first step in making accurate diagnoses is to assess the quality of the imaging. Patient positioning is critical to ensuring a valid diagnostic result, and radiologists are often faced with the task of screening X-rays for positioning or other image acquisition errors.

With the implementation of AI, automation of quality control in MSK CR images could be accomplished at the point of imaging, when the patient is still available and corrected images could be acquired. For example, AP standing pelvic X-rays require specific positioning of the lower extremities and hips. Errors in positioning can lead the radiologist or orthopedic surgeon to order a new X-ray, which costs time and effort for the clinician and patient.

As we know, deep learning requires large amounts of annotated image data for training models to be robust. In the pelvic X-ray quality

control example, training data should encompass a range of potential positioning errors from a variety of patient groups. No amount of training data will be able to completely reproduce the scope of the problem, but the robustness of the solution will benefit from the diversity of pelvic representations in the training examples.

Anomaly Detection

Clinically relevant anomalous findings in MSK X-ray include fractures, tumors, lesions, and bone marrow edemas, to name a few. Anomaly detection is perhaps the most challenging task for AI because they may occur anywhere within the skeletal tissue and to varying degrees. In contrast, anatomy-specific findings such as hip dysplasia, knee osteoarthritis, or scoliosis can be more readily modeled using AI given a suitable amount and variety of training images.

Beyond the technical challenges of anomaly detection are the clinical implications for erroneous findings. False-positive detection of a bone tumor, for example, could lead to unnecessary follow-ups and highly invasive diagnostic procedures which affect the patient's well-being and cost the healthcare system time and resources.

False negative, or missed findings, pose even greater risks to both patients and practitioners. A missed spinal fracture diagnosis in an osteopenic elderly patient, or a missed tumor in a pediatric patient, could lead to diminished quality of life and untimely death, along with the legal ramifications for the clinicians and institutions, although we fully acknowledge that the risk to patients is the first concern in designing any diagnostic tool.

IB Lab KOALA: Assessment of the Gonarthrosis Stage

Knee osteoarthritis is a painful and immobilizing joint disease that can lead to joint replacement. Knee osteoarthritis has a lifetime risk of up to 45% [2], with risk determined based on two main risk factors: aging and obesity [2, 3]. Knee osteoarthritis affects over 200 million patients worldwide

[4], resulting in approximately 100 million knee radiographs in the EU alone in 2020. Consistent tracking of radiographic changes over time could aid in early detection and prevention of disease progression. Due to the semiquantitative nature of the score, the individual diagnosis in the Kellgren-Lawrence score shows very low inter- and intraindividual agreement and is thus unfavorable for standardized treatment decisions, as well as prerequisites for efficacy studies of osteoarthritis therapies. Radiologists read an average of ten knee radiographs per day, which corresponds to approximately 40 min of daily workload.

Deep learning algorithms were trained on over 35,000 individual knee radiographs containing data from a longitudinal study of centers in the United States. Each image was evaluated by board-certified radiologists using the OARSI criteria and the Kellgren-Lawrence consensus scales.

The AI-driven KOALA software evaluates the stage of osteoarthritis according to the Kellgren & Lawrence grading system, is validated on over 10,000 knees, and is available as a CE-marked or an FDA cleared version. KOALA also provides accurate and automated measurements of minimum joint space width along with assessment of severity of joint space narrowing, presence of osteophytes and sclerosis based on OARSI criteria. Results are summarized in visual output reports, attached to the original radiograph and automatically stored in the PACS system. KOALA facilitates monitoring of disease progression by allowing comparison of radiographic disease parameters over time. The results must be confirmed or rejected by the assessing physician and is thus an assistance system for improving the efficiency and standardization of radiographs in knee osteoarthritis (Fig. 2).

IB Lab PANDA: Determination of Pediatric Bone Age

PANDA is used to assess bone age for predicting height in children. PANDA uses an ensemble of decision models to determine bone age based on the Greulich and Pyle atlas. The standard



Fig. 2 KOALA knee osteoarthritis report and analysis results displayed on the screen

derivation for a given chronological age is determined by rounding down to the nearest age in the Brush table for the corresponding sex [5].

PANDA was trained on over 12,000 hand radiographs from two institutions in the United States (Lucile Packard Children's Hospital of Stanford University and Children's Hospital Colorado) [6]. Validation was performed using the RSNA 2017 Bone Age Challenge test set. Each image in this test set was reviewed by three radiologists trained in pediatrics. IB Lab's diagnostic support tool, PANDA, uses deep learning technology to report bone age based on the Greulich and Pyle scale, saving time by presenting results within 5 s. PANDA's automated measurement of bone age is accurate to the mean absolute deviation of 4.3 months as measured on the test set from [7]. PANDA provides accurate data for decision-making in growth-associated disorders, and standardized measurements and reporting schemes facilitate monitoring of treatment progress. Manual estimation by comparing digital radiographs to reference images in the Greulich and Pyle atlas is tedious and has a high degree of variability between readers. PANDA provides a rapid automated method for estimating bone age and

monitoring growth and development in children (Fig. 3).

IB Lab HIPPO: Measurement of the Hip and Pelvis

HIPPO is used for measuring hip positioning and pelvic morphology by measuring the most common angles and distances on an X-ray of the pelvis. The HIPPO module is intended for adults between 18 and 90 years of age with hip pain, suspected congenital disease, femoral-acetabular impingement, or osteoarthritis of the hip.

HIPPO was developed using deep learning algorithms trained on over 4000 individual radiographs of the pelvis and hip. It reproducibly detects and localizes anatomically relevant landmarks on the hip and pelvis (e.g., tear figure) automatically. The AI follows the established radiological workflow: measurement of anatomical distances and angles, detection of disease morphologies, and standardized classification and reporting. There is a consensus assessment of each radiograph: Each detection step is performed by an ensemble of three

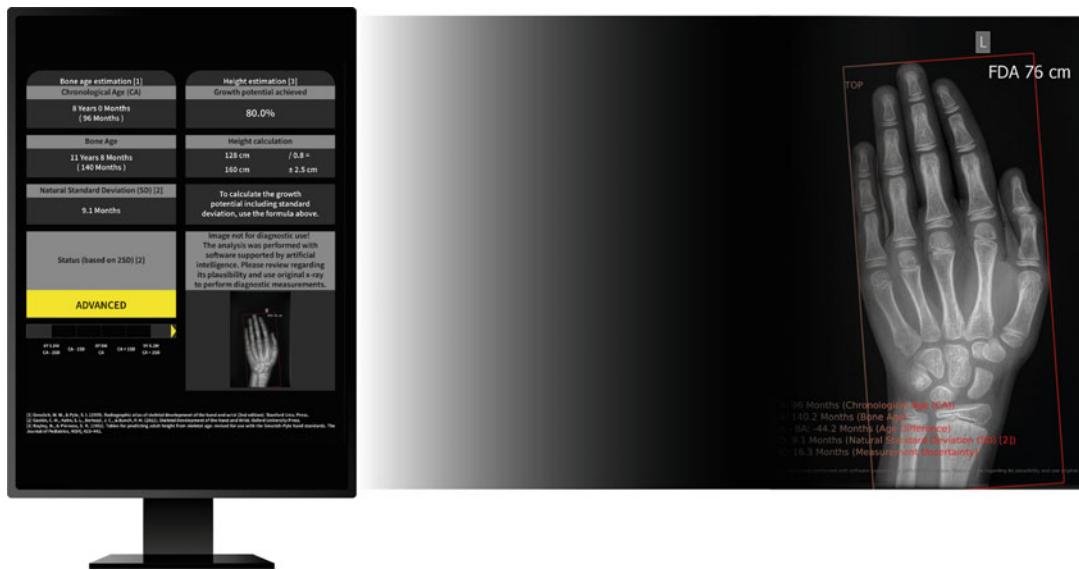


Fig. 3 PANDA bone age report and analysis results visible on the screen

AI models that vote for the best result and increased precision.

Using hip angle measurements, physicians can take appropriate action and therapy for early signs of hip disease, including arthritis and dysplasia. HIPPO performs objective, standardized measurements of key hip angles on digital radiographs. These include CCD and LCE angles, as well as Tönnis angle (acetabular index), Sharp's angle, and femoral extrusion index. In bilateral standing AP hip radiographs, HIPPO assists the medical expert in detecting the presence or absence of leg length discrepancies.

Femoroacetabular impingement and hip dysplasia are the two main causes of secondary hip degeneration, leading to hip replacement in the final stage. Hip arthroplasty is projected to increase from 1.8 million per year in 2015 to 2.8 million per year in 2050 in OECD countries [8], resulting in an increased workload for radiologists. Reading pelvis radiographs requires thorough knowledge of 3D pelvic morphology to correctly interpret the 2D projection; this is difficult, subjective, and prone to error. Unstructured reporting further leads to inconsistencies in the diagnostic process. Radiologists read an average of ten hip radiographs

per day, representing approximately 40 min of daily workload [9] (Fig. 4).

IB Lab LAMA: Measurement of the Whole Leg

LAMA is used to measure leg length discrepancies and detect deformities of the knee axes and can be used in the decision to perform an osteotomy.

LAMA is a fully automated radiological imaging software designed to help medical professionals to measure leg axis geometry. LAMA helps detect knee axis alignment deformities by performing the following measurements: mechanical Lateral Proximal Femoral Angle (mLPFA), mechanical Lateral Distal Femoral Angle (mLDFA), mechanical Medial Proximal Tibial Angle (mMPTA), mechanical Lateral Distal Tibial Angle (mLDTA), Mechanical Axis Deviation (MAD), Hip Knee Angle (HKA), and Anatomic Mechanical Angle (AMA) in standing AP radiographs of the whole leg. LAMA helps identify leg length discrepancies by providing the following measurements: femur, tibia, and full leg length, as

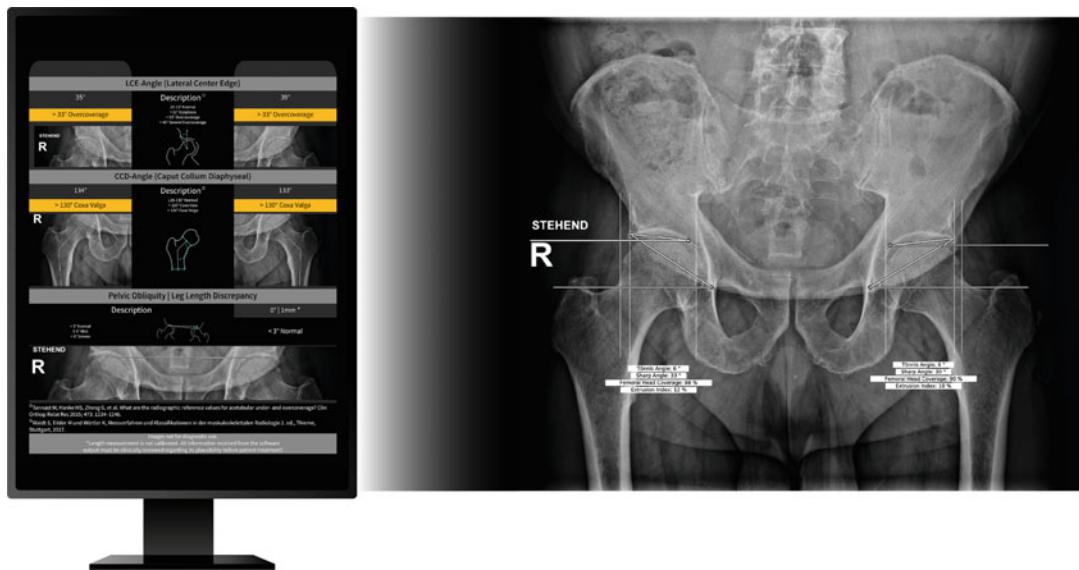


Fig. 4 HIPPO hip/pelvis report and analysis results visible on the screen

well as the difference between right and left legs on bilateral images. Lower limb leg or length discrepancies are common deformities that affect most adults and children; LAMA increases workflow efficiency by saving reading and reporting time.

Lower extremity leg or length discrepancies are common deformities affecting up to 70% of the adult and pediatric population [10]. When undetected or inaccurately measured, patients suffer functional and biomechanical limitations as well as cosmetic impairment. Even minor deviations can lead to imbalances and unilateral pain throughout the body and trigger passive structural and degenerative changes in the hip, spine, knee, and muscles. Accurate, reliable measurements of lower extremity geometry require expert training and specialized software in established protocols. Standardization through reproducible AI that automatically measures at the expert level can help avoid errors and train junior doctors. The diagnostic support tool LAMA uses deep learning technology to automatically and accurately measure leg geometry to evaluate lower extremity deformities. LAMA helps detect genu varum/valgum by measuring mechanical axis deviation (MAD) and detect leg length discrepancies by comparing the right and left legs on bilateral

images. AI facilitates monitoring of disease progression by facilitating comparison of radiological disease parameters over time (Fig. 5).

AI Disrupts Status Quo of Radiology Reporting

Status quo: nonstructured, manual workflows

Currently, the existing workflows for diagnostic and reporting of disease findings are non-structured and manual. Disease symptoms are often recognized at stages either too late to apply the appropriate treatment (e.g., OA), or the diagnostic results are based on nonstandardized radiological parameters (e.g., for pre-/postsurgery assessments). The agreement rate in terms of disease classification among physicians can be as low as 30%. We aim to change this by providing a continuous workflow to enhance the diagnostic and reporting workflow with standardized and accurate disease findings.

We change managements: augmenting the workflow with more precise and novel imaging biomarkers for new treatments and proactive intervention.

We disrupt the status quo by empowering physicians and scientists with powerful tools to



Fig. 5 Whole-leg report and analysis results visible on the screen

assess, track, and predict the disease. Our software offers an augmentation and standardization of the highly subjective (and time-consuming) assessment process of 2D/3D imaging in MSK radiology. The MSK platform analyzes data in near real time, whereas physicians need minutes to look through and assess and document the status of a scan. For this purpose, we build upon our existing know-how in image analysis and interpretation. In our case, the application of deep learning is recognition of MSK-relevant scoring and measuring tasks in radiological images. We enrich the findings with new bone health imaging biomarkers, such as insights into the bone micro architecture. Therefore, our objective is the detection and digital documentation of clinically relevant features in MSK imaging data beyond what can be performed by human readers in the course of a regular working time.

Take-Home Message: Artificial Intelligence (AI) in Musculoskeletal Radiology

- AI relies on deep learning to automatically detect and localize anatomically relevant landmarks at the hip, knee, and ankle.

- AI improves the diagnostic accuracy of knee osteoarthritis according to the latest clinical guidelines.
- AI facilitates (radiological) monitoring of the progression of various bone diseases such as osteoarthritis of the knee.
- AI enables early radiological signs to be detected and allows conclusions to be drawn about (possible) prognosis.
- AI increases patient throughput by automating the report generation process.
- AI reduces the impact of inter-rater variability in radiographic morphology assessment.

References

1. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC Med Res Methodol*. 2019;19:64.
2. Murphy L, et al. Lifetime risk of symptomatic knee osteoarthritis. *Arthritis Care Res*. 2008;59:1207.
3. Losina E, et al. Lifetime risk and age of diagnosis of symptomatic knee osteoarthritis in the US. *Arthritis Care Res*. 2013;65:703.
4. GBD 2015 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the global burden of disease study 2015. *Lancet*. 2016;388:1545.

5. Greulich WW, Sarah IP. Radiographic atlas of skeletal development of the hand and wrist, 2nd edition. Stanford University Press, 1959.
6. Larson DB, Matthew CC, Matthew PL, Safwan SH, Nicholas VS, Curtis PL. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. *Radiology*. 2018;287(1):313–22. <https://doi.org/10.1148/radiol.2017170236>
7. Halabi SS, Prevedello LM, Kalpathy-Cramer J, Mamounov AB, Bilbily A, Cicero M, ... Flanders AE. The RSNA pediatric bone age machine learning challenge. *Radiology*. 2018;290(2):498–503.
8. Pabinger C, Lothaller H, Portner N, Geissler A. Projections of hip arthroplasty in OECD countries up to 2050. *Hip Int*. 2018;28:498.
9. Carlisle J, Zebala LP, Shia DS, Hunt D, Morgan PM, Prather H, Wright RW, Steger-May K, Clohisy JC. Reliability of various observers in determining common radiographic parameters of adult hip structural anatomy. *Iowa Orthop J*. 2011;31:52–8.
10. Knutson GA. Anatomic and functional leg-length inequality: a review and recommendation for clinical decision-making. Part I, anatomic leg-length inequality: prevalence, magnitude, effects and clinical significance. *Chiropr Osteopat*. 2005;13:11.



AIM and Explainable Methods in Medical Imaging and Diagnostics

35

Syed Muhammad Anwar

Contents

Introduction	502
Medical Imaging in Clinical Diagnosis	502
Recent Advances in Technology Toward Medical Imaging	503
Machine Learning-Based Methods	504
Explainable Artificial Intelligence: XAI	505
Transparent Models	506
Post Hoc Interpretability	507
Explainable Models for Deep Neural Networks	507
Conclusions and Future Landscape for ML-Based Imaging and Diagnostic Methods	508
References	509

Abstract

Medical imaging and diagnostics have benefited from recent advances in machine learning in general and deep learning in particular. There are a large number of studies that report significant performance in a variety of medical image analysis tasks. At the same time, advances in time series analysis of electronic health records and physiological signals from a variety of sources (including clinical and nonclinical

wearable sensors) have enabled efficient clinical pipelines and diagnostic performance. While there are a large number of studies that report such advances and use cases, there is a certain level of resistance to clinical adaptability of such methods. We attribute one of the major reasons for this to be the black-box nature of such machine learning and data-driven models. While clinical decision-making, for example, using radiology, relies not only on radiographic information but also on clinical expertise (that could rely on clinical history of a patient as well as demographic information), such informed decision makes the backbone of clinical practice. Hence, machine learning should adapt to making such informed decisions for wider

S. M. Anwar (✉)

Department of Software Engineering, University of
Engineering and Technology, Taxila, Pakistan

adaptability and finding more practical use cases. We argue that explainable methods, where a machine learning-based model provides supporting information for their decisions making them understandable, will be more representative of what is required in clinical practice. In turn, such methods would augment the current medical imaging and diagnosis pipeline and hence allow adaptation of artificial intelligence in clinical workflows. This would transform the healthcare industry with far-reaching benefits for both clinical practitioners and patients enabling precise and personalized medicine.

Keywords

Deep learning · Machine learning · Medical images · MRI · CT · PET · Diagnosis · Explainability · Generalized models · Artificial intelligence

Introduction

In recent years, we have witnessed an explosive growth of applications utilizing artificial intelligence (AI) to solve a wide array of technological problems. This wave of recent success for AI-based methods is driven by the advancements in both hardware and software platforms. Thereby, AI and machine learning (ML) are adapted as technology of choice in areas which were not considered apt for such advancements a few years back. Among applications are included areas such as cybersecurity, social media, medical industry, and industry 4.0 transformation. The success of AI is fuelled by the availability of more compute power, large-scale datasets, and optimized computing libraries. In particular for medical industry, the revolution is just around the corner as we see more and more clinical tasks being affected by AI algorithms. A driving force has been the widespread digitization, resulting in the availability of big data. We now have hospitals having large-scale electronic health records storing patient information, radiographs, lab exams, and all sorts of information. A machine

learning algorithm will take these numbers as variables and attempts to learn a model that can predict patterns within the data.

There are certain challenges in this field: i) the amount of data for particular scenarios which could be limited, ii) the quality of the labels assigned to data, and iii) the ability of clinical experts to understand what the machine learning algorithm is predicting. While large-scale digitization in clinical settings means that big data is available, privacy and security concerns do not allow this data to be freely used and hence introduce certain boundaries. One of the reasons for these security concerns arises from the fact that it is not exactly clear how machine learning models are going to use this data. The amount of research coming out in this field and the level of interest shown by industry to roll out ML-enabled products for clinical pipelines are a clear indication of the fact that clinical practice in the twenty-first-century is going to change drastically. Herein, we present an overview of explainable artificial intelligence (XAI) which is believed to solve many problems associated with the adoption of AI in clinical pipelines. Since it is a developing field, we aim to identify the terminologies used and discuss how XAI is important in medical imaging and diagnosis. We highlight some of the methods that are being successfully used and also comment on the future direction this field could take. Hence a comprehensive overview of existing methods and taxonomies is presented.

Medical Imaging in Clinical Diagnosis

Medical imaging is a technique used for noninvasively looking inside the human body. This is used for both diagnosis and treatment planning. Moreover, medical imaging aids in early detection of a clinical conditions. There are different modalities used in the field of medical imaging with a wide variety of clinical applications. The use of imaging in clinical diagnosis and intervention planning has been prevalent for some time now. However, the use of automated image analysis techniques in recent years has resulted in an increase in the importance and significance of imaging in clinical

diagnostics. In addition, these imaging methods allow advances in medical research and have resulted in discovery of new knowledge and understanding of working of various body organs. Moreover, the progression of certain medical conditions is now better understood due to studies involving medical images acquired during different stages of the disease [1]. Hence, we can safely conclude that medical imaging has benefited the field of diagnosis, prognosis, and clinical research.

Most commonly used radiographic imaging techniques include X-rays, magnetic resonance imaging (MRI), computed tomography (CT), positron-emission tomography (PET), ultrasound, and mammography. Among these modalities, X-rays have been widely used for a long time due to their large-scale availability and lower cost. In terms of safety, there are certain concerns with X-ray-based imaging modalities (including CT). This is where MRI has been largely successful in terms of both safety and image quality. Recent MR scanners are versatile in terms of the body organs and the level of contrast (for both hard and soft tissues) and hence result in capturing a major stake in radiographs currently generated. Similarly, ultrasound generates images using sound waves and hence is considered safe. Ultrasounds are used for diagnosis of a variety clinical conditions in situations including vascular diseases and pediatrics. Some areas of medical image analysis have seen far better advances with human-level performance, such as in areas related to lung diseases [2] and lesion detection in mammograms [3]. Although the predictive performance of these methods is significant, the clinical community still demands further investigation before these are adopted in practice. Moreover, certain areas are still struggling such as pancreatic tumor detection, and hence innovation in technology and methods is required for clinical-level performance.

Recent Advances in Technology Toward Medical Imaging

There are certain areas of technology that have seen significant advancements in terms of their utility in medical imaging and diagnostics. This

has enabled machines to perform very complex tasks with human-level performance in fields such as medicine. Among these areas, AI is leading the way toward working with radiologists and revamping the clinical pipelines for better outcomes for patients and clinical diagnosis. Hence, AI-enabled technology will provide a holistic approach to medical systems. This would include AI-enabled guidelines for doctors to select a particular test, instead of performing different tests before coming up with the right one. In terms of acquiring a radiograph, AI would enable better image reconstruction strategies and help in identifying acquisition parameters that are better suited to the task at hand. Further, AI could help in dose optimization for generating medical images as well as image quality enhancements post acquisition (allowing to capture images with a lower level of dose). For reporting, AI and ML is already starting to augment radiologist to ensure attention to details and better and faster reporting (using segmentation and detection algorithms). AI is also leading the way in developing solutions for personalized treatment with precision medicine and allowing better optimization in utilizing the currently available healthcare facilities. While the industry is on this path of realizing AI-enabled solutions in all aspects of healthcare systems, there are certain limitation to currently existing methods and algorithms. Among these, one of the major considerations is the development of explainable AI. It is required to ensure that the models used in critical applications, such as medicine, are better understood and hence allow making more informed decisions.

Other than AI, some of the leading technological advances that are impacting medicine include the development of augment reality-based (AR) systems, utilization of three-dimensional printing, and the digital twin technology. One of the major beneficiaries from the development of AR systems is how surgeries are performed. In particular, using 3D imaging services and AR-based systems, surgeons are being empowered to make better surgery planning and decisions. Using 3D printing also helps in explaining the surgical procedures to patients and could allow better understanding in clinical

research. Hence a combination of 3D printing and AR would enable improved surgical decisions and allow patients to better understand surgical options and make more informed choices. For these methods, explainability is again important when machine learning algorithms are used for designing such systems. AI is also empowering the digital twin technology which would also benefit from explainable systems. The digital twin technology would allow clinical experts to allow best therapy solutions by testing it on the digital twin before administering it to a patient and hence achieve best possible outcomes for patients.

Machine Learning-Based Methods

To put it in simple terms, machine learning is going to disrupt how medicine is practiced [4]. Current medical systems are rule-based system, governed by a set of defined rules. But in the coming days, clinicians are going to rely on rules that are learned by the data themselves. Hence for the machine learning models to perform in an acceptable manner, the quality and quantity of the data for these “data-hungry” models is the most important factor. Another challenge of these data-driven models is that they do not guarantee a causal inference. This is where explainable models play a key role.

Medical image analysis has been a major beneficiary of the current developments in artificial intelligence. While the predictive performance of the models is on the rise, there are certain limitations of these methods which remain as a bottleneck in clinical adaptation of these new techniques. The advancements have been supported by the availability of large-scale datasets for a wide variety of applications. Data are now available for various imaging modalities such as magnetic resonance imaging, computed tomography, and X-rays. For instance, multimodal brain tumor segmentation (BRATS) [5] data are a widely used benchmark for brain tumor segmentation using MRI images and machine learning methods. A large collection of data, fastMRI [6], are available for MR image reconstruction studies. In particular raw data from single- and multi-coil acquisition scenarios

are publicly available enabling the development of ML models for this important area. It should be noted the deep learning (DL) models continue to outperform linear models and shallow networks in all forms of medical image analysis tasks [7].

In [8], a convolutional network was used to develop a model for retrieving medical images. Since the size of databases containing radiological images continues to increase all over the world, such a method is important for retrieving the most relevant radiographs. This would enable better diagnosis and prognosis and allow clinicians to study relevant patients before making diagnosis for new patients. In [9], a deep learning model was presented to segment brain tumor with considerable accuracy using the BRATS dataset. Recently, the focus has been on survival prediction for tumor subjects, where a precise segmentation of tumor regions plays a critical role [10]. While for explainable decision-making, radiomics features are preferred [11], but with limited data the predictive performance of such models can be affected. In [12], deep learning is used for 3D input MR images and tumor regions are segmented. Further, radiomics features are extracted from the tumor regions for survival prediction task, where decision trees are used for making the prediction. Machine learning has also been widely successful in many other segmentations and diagnosis tasks such as Alzheimer disease detection [13, 14] and multi-tissue segmentation [15]. It is also shown that combining multimodal data can improve the performance in clinical diagnosis [16]. Data-driven approaches using machine learning are also contributing to the field of image reconstruction [6]. It has been shown that deep learning can be used to optimize the subsampling patterns for MRI acquisition [17].

Although for medical image analysis, deep learning has been found to be successful, there are certain clinical application where the application of such methods is limited. One of the biggest challenges has been to use DL in scenarios where the data is limited, e.g., in electroencephalography (EEG) [18] and electrocorticography (ECoG) [19]. It should be noted that such transparent methods are better in terms of understandability, but their predictive power is limited. In another line of research, such models have been used to

generate information by learning knowledge from another imaging modality [20]. Machine learning has also been used for predicting patient outcomes by learning from 1D and low-dimensional data [21, 22]. In summary, there is no question on the application and usability of ML-based models in medical imaging and clinical diagnosis when we consider their predictive power. But the challenges arise from the black-box nature of such models and in cases where the reasons for the model coming up with certain predictions are not understandable by the stakeholders involved. In the next section, we provide an overview of the field of explainable AI and how this is important for clinical practice.

Explainable Artificial Intelligence: XAI

For critical areas of application such as medicine, machine learning models face a significant challenge when opaque models are used. Explainable methods are the foundation of explainable artificial intelligence – XAI, also called responsible AI in some cases. A high-level difference between ML and XAI is presented in Fig. 1, where the major difference is the introduction of a black box that enables explanation for the end user in terms of why a model makes a certain prediction. There are a wide range of definitions prevalent in literature for XAI, but no single definition has been agreed upon by all stakeholders yet. Moreover,

there are different terminologies used in literature which directly or indirectly refer to explainable methods. Interpretability is one of the most common terms that is used interchangeably with explainability, but it needs to be understood that both these terms represent different phenomenon. In particular, interpretability is a passive property of a model and broadly represents how much a model makes sense to a human while using that particular model. This could mean how much the person understands in terms of how the model is set up and working. On the other hand, explainable methods have a property where the model demonstrates the characteristics of clarifying its internal functions. In general, a model is considered to be understandable, if humans can understand the function of that model without going into the details of the internal functions or algorithms used. Hence, incorporating interpretability to a model means using methods to provide a meaning for humans in certain humanly understandable terms. Those methods that are interpretable are considered to be transparent, since the decisions made by such methods are understandable by humans – the key stakeholders in high-risk applications.

There are two aspects of explainable models in terms of understanding including model understanding and human understanding. Thereby the role of the audience (end user) is very important and must be considered while developing such methods. The audience could be diverse and

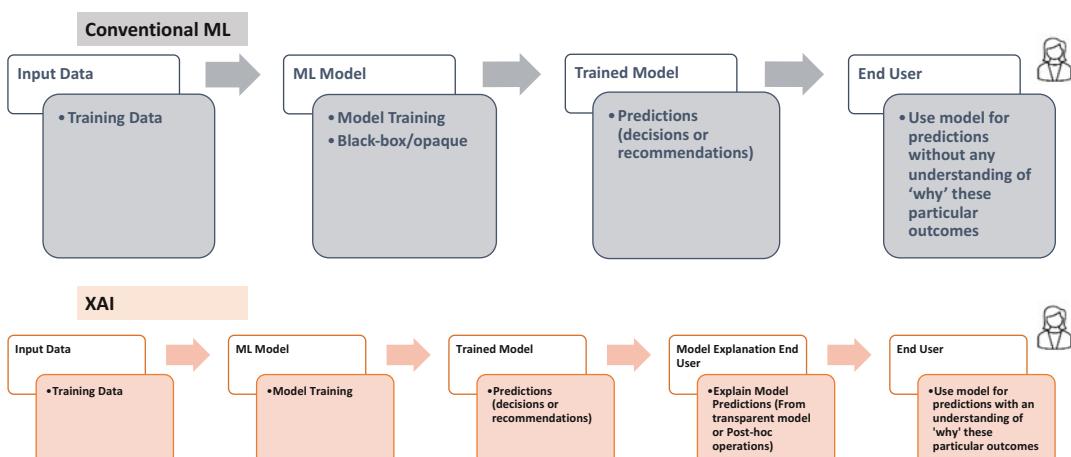


Fig. 1 Difference between conventional machine learning and explainable artificial intelligence pipelines

depends on the application including domain experts, data scientists and ML practitioners, regulatory bodies, executives, and end users. Hence the interpretations could vary depending on the type of audience. Moreover, this has also resulted in various goals that are being pursued in the field of XAI. One of the major goals of XAI is to achieve trustworthiness – a measure of confidence that the model will achieve the expected outcome in a certain situation. For medical applications, trustworthiness is critical, since both medical practitioners and end users are highly concerned on what AI would mean for them. A trustworthy model is critical in clinical decision-making. The goals of robustness and generalizability are also significant in medical domain. Hence explainable models need to have an associated level of confidence to the decision that these methods make. The models also need to be transferable and hence allow application in areas where data is limited. Transferability also accounts for the fact that a trained model can be successfully applied to the test (never before seen) data.

Another major concern in using AI in real-life application of critical nature is the ethical considerations. Explainable methods aim to achieve fairness and hence ensure that AI is used in an ethical and fair manner. For clinical applications, where human life could be at stake, it is important to achieve fairness in the decisions the AI-based models will make. To make models more interpretable and explainable, humans in the loop methods are being developed [23]. This ensures that humans are included in various stages of the AI-based decision-making. For example, interactive models are developed that learn from the human interaction with medical images, while radiologists make their readings in real time [24]. This data is used for training models that are expected to perform better than those models that involve no human interaction. Human in the loop also means that humans should be able to interact with ML models and the outcomes should be accessible.

Another important aspect in explainability is to define causability [25], which measures what the explainable model achieves in terms of causal understanding for a human being. Here causal relation defines the cause-and-effect relationship

[26]. It is imperative for explainable models, particularly in the medical imaging and diagnosis domain, to achieve a certain level of causability understanding. One way of achieving this is to fuse information from different modalities. In particular for use of AI in clinical practices, data is derived from diverse sources including imaging (MRI, PET, CT, etc.), clinical history, personal demographic information, lab tests, and physiological signals to name a few. Generating multi-modal embeddings in an explainable way could be the way forward to generate precise and personalized diagnosis. For generating such embedding arising from cross-modal data fusion, the semantic gap that exists between these varying data sources needs to be overcome. One way to achieve such interpretable representation is to use graph representation learning [25] which, although a promising direction, requires significant research.

Explainable methods are broadly categorized in two types: i) transparent models and ii) models that require post hoc operations. This categorization stems from the fact that some ML models are inherently simple and hence interpretable, in terms of both model understanding and human understanding, while complex models, such as deep neural networks which are not comprehensible for human, require post hoc operations that enable interpretability and explanations for the decisions the model makes.

Transparent Models

Such models are simple enough to be explainable by themselves and include classical machine learning algorithms such as logistic regression and K-nearest neighbors. While such models are simple and hence comprehensible for a human observer or domain expert, these models could be limited in terms of their predictive power. Moreover, the transparency of such models could be lost as soon as more data or complexity is added to such models. One of the reasons for success of more complex models, such as deep learning, has been their higher predictive capability even when large-scale data is involved. However, the added complexity is also the reason for

such model to be considered as opaque or black box, since they are no more interpretable by themselves. There are other ML methods that are considered to be transparent such as decision trees and simple rule-based learners. In particular, rule-based learners have a very intuitive nature and hence also led to the success of AI in a wide array of application before the advent of deep learning models. While such algorithms have some nice explainable properties, but for a lot of real-world application, defining a robust set of rules covering all seen and unseen scenarios has been a challenge. The models developed using probabilistic basis are also considered to be transparent, such as Bayesian models. Hence there has been a wide interest in developing such models even when deep learning is used for better interpretability and explainability. Inference from such models could lead toward identifying the causality as well as estimating the uncertainty associated with decisions allowing to build confidence in the decision taken by ML-based methods.

Post Hoc Interpretability

For ML models that are not transparent, methods are adopted that can explain how the output is produced by a trained model. The aim for these post hoc techniques is to communicate an understanding of the predictions generated by the ML model. These post hoc methods are further categorized based on the type of ML models these are used for. Model agnostic techniques are designed to be independent of the type of underlying model used. Such techniques include visualization methods, identification of feature-level importance, and model distillation to simplify models and hence understand the predictions. Moreover, for shallow ML models, such as support vector machines and ensemble techniques, feature relevance identification and visualization methods are used as post hoc operations. For deep learning, post hoc operations could involve model simplification, for which there are some standard methods presented in literature. Since a lot of advances in medical imaging and diagnosis are driven by adopting deep learning models and these models

are inherently opaque, we particularly look at some standard methods (performing post hoc explanations) for such models.

Explainable Models for Deep Neural Networks

Deep neural networks, by virtue of having a large parameter space and layers, have an inherently predictive performance when compared with linear and shallow ML models. There is a wide array of methods presented in literature for introducing post hoc explainability when using deep neural networks. We present a summary of four main areas here as reference.

In some methods, a simple transparent model (surrogate) is used at the output replacing the decision function. An example of such method is the local interpretable model agnostic explanation (LIME) algorithm [27]. The algorithm incorporates decision trees and introduces sparsity in an attempt to model surrogate linear explainable models. For analyzing feature-level importance, occlusion analysis can be performed [28]. In this method different patches or features are occluded, and the model output is analyzed. One way to analyze the output is to create heatmaps and hence identify which parts of the image or features are most affected. In Shapley (SHAP) analysis [29, 30], a larger set of features are occluded and hence have a global level of understanding. With deeper neural networks, the problem of shattered gradient arises, where the local gradient can look like noise and the changes in gradient over different layers are also significant. Explanations based on these gradients can look like noise and hence are not useful. SmoothGrad [31] is a technique used to overcome the shattered gradient phenomenon and hence can aid in model explanation for deeper neural networks. Generally, deep neural network follows a feed-forward cycle followed by back propagation while training. Inspired by this learning mechanism, layer-wise relevance propagation (LRP) [32, 33] mechanism is used to introduce explanations in the process. The LRP is a forward-backward loop where the relevance values are propagated through the neural

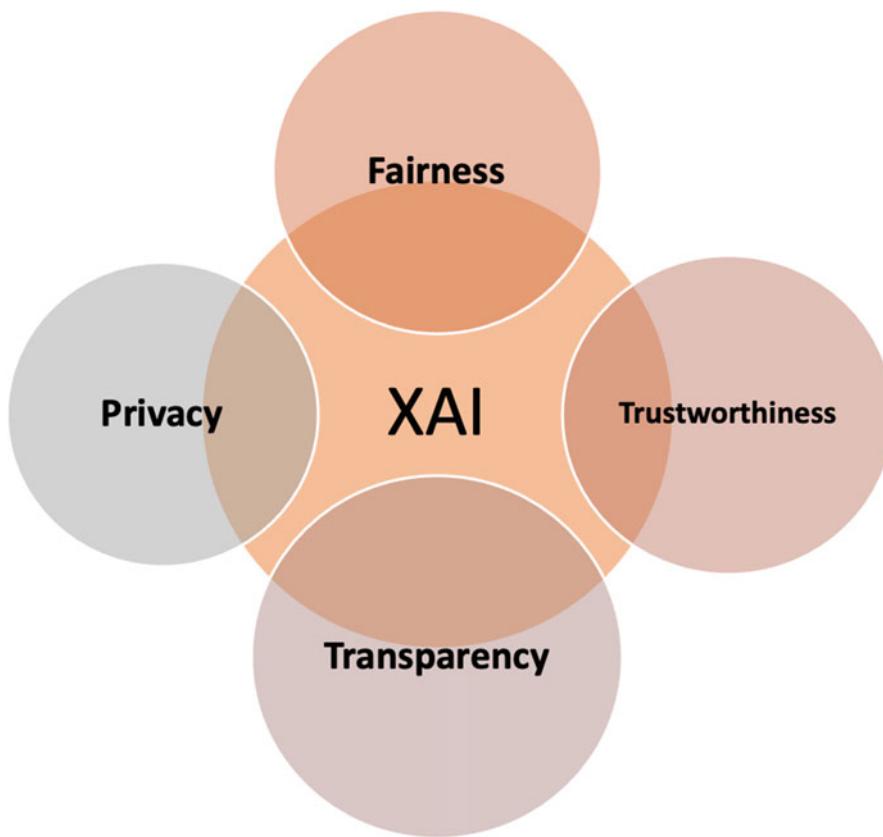


Fig. 2 The goals that explainable artificial intelligence aims to achieve with responsible AI

network from the input to the output and back and in the process generate explanations. There are other methods present in literature driven from these basic models to incorporate post hoc explanations in deep networks and can be found here [34]. In general, the kind of things expected from XAI is presented in Fig. 2. It is expected that XAI will enable trustworthiness, fairness, privacy, and transparency and lay the foundations for responsible AI.

Conclusions and Future Landscape for ML-Based Imaging and Diagnostic Methods

Over the last decade, we have seen a resurgence in the pace with which AI has been deployed in real-world application including medicine. The future lies in precision and personalized medicine, enabling better outcomes for all

stakeholders benefitting from the digitization and technological advancements. While this will be realized once the opaque (black-box) nature of currently used ML methods is transformed into more interpretable and explainable models, one emerging area which is closely related to explainability is the assessment of privacy while using ML and AI in clinical pipelines. Current data-driven models are known to memorize information, and with their black-box nature, it is not clearly understood what kind of information they could store. While the aim of current research is to develop models with high predictivity, with interpretation we cannot fully understand what kind of information these models hold and how this can be used by various stakeholders. Hence XAI would need to focus on privacy awareness to ensure that the critical information is secure, and it is better understood that in what ways the internal relations learned by a trained model

can be interpreted by those that have access to this information.

In conclusion, as the field of machine learning is progressing at a fast pace and in the process revolutionizing medical imaging and diagnosis, it is also creating new avenues and challenges. XAI is at the forefront of overcoming most of these challenges. With explainable AI, we can hope to achieve responsible AI, and this would enable a large-scale adaption of AI in clinical pipeline with more confident and satisfied stakeholders. The field is wide open and required significant research effort and innovation. One thing is for sure, that with XAI, medical practice will benefit in the coming days allowing improved diagnosis and better patient outcomes.

References

1. Feng Z, et al. Early prediction of disease progression in COVID-19 pneumonia patients with chest CT and clinical characteristics. *Nat Commun.* 2020;11(1):1–9.
2. Li X, et al. Multi-resolution convolutional networks for chest X-ray radiograph based lung nodule detection. *Artif Intell Med.* 2020;103:101744.
3. Rodriguez-Ruiz A, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *JNCI: J Natl Cancer Inst.* 2019;111(9):916–22.
4. Obermeyer Z, Emanuel EJ. Predicting the future – big data, machine learning, and clinical medicine. *N Engl J Med.* 2016;375(13):1216.
5. Menze BH, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imaging.* 2014;34(10):1993–2024.
6. Zbontar J, et al. fastMRI: an open dataset and benchmarks for accelerated MRI. *arXiv preprint arXiv:1811.08839.* 2018.
7. Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK. Medical image analysis using convolutional neural networks: a review. *J Med Syst.* 2018;42(11):1–13.
8. Qayyum A, Anwar SM, Awais M, Majid M. Medical image retrieval using deep convolutional neural network. *Neurocomputing.* 2017;266:8–20.
9. Hussain S, Anwar SM, Majid M. Segmentation of glioma tumors in brain using deep convolutional neural network. *Neurocomputing.* 2018;282:248–61.
10. Yousaf S, RaviPrakash H, Anwar SM, Sohail N, Bagci U. State-of-the-art in brain tumor segmentation and current challenges. In: *Machine learning in clinical neuroimaging and radiogenomics in neuro-oncology.* Springer; 2020. p. 189–98.
11. Yousaf S, Anwar SM, RaviPrakash H, Bagci U. Brain tumor survival prediction using radiomics features. In: *Machine learning in clinical neuroimaging and radiogenomics in neuro-oncology.* Springer; 2020. p. 284–93.
12. Sun L, Zhang S, Chen H, Luo L. Brain tumor segmentation and survival prediction using multimodal MRI scans with deep learning. *Front Neurosci.* 2019;13:810.
13. Farooq A, Anwar S, Awais M, Rehman S. A deep CNN based multi-class classification of Alzheimer’s disease using MRI. In: *2017 IEEE international conference on imaging systems and techniques (IST).* IEEE; 2017. p. 1–6.
14. Jo T, Nho K, Saykin AJ. Deep learning in Alzheimer’s disease: diagnostic classification and prognostic prediction using neuroimaging data. *Front Aging Neurosci.* 2019;11:220.
15. Anwar SM, et al. Semi-supervised deep learning for multi-tissue segmentation from multi-contrast MRI. *J Signal Process Syst.* 2020;1–14.
16. Altaf T, Anwar SM, Gul N, Majeed MN, Majid M. Multi-class Alzheimer’s disease classification using image and clinical features. *Biomed Signal Process Control.* 2018;43:64–74.
17. Bahadir CD, Wang AQ, Dalca AV, Sabuncu MR. Deep-learning-based optimization of the under-sampling pattern in MRI. *IEEE Trans Comput Imaging.* 2020;6: 1139–52.
18. Saeed SMU, Anwar SM, Majid M, Bhatti AM. Psychological stress measurement using low cost single channel EEG headset. In: *2015 IEEE international symposium on signal processing and information technology (ISSPIT).* IEEE; 2015. p. 581–5.
19. RaviPrakash H, et al. Deep learning provides exceptional accuracy to ECoG-based functional language mapping for epilepsy surgery. *Front Neurosci.* 2020;14:409.
20. Masoudi S, Anwar SM, Harmon SA, Choyke PL, Turkbey B, Bagci U. Adipose tissue segmentation in unlabeled abdomen MRI using cross modality domain adaptation. In: *2020 42nd annual international conference of the IEEE engineering in medicine & biology society (EMBC).* IEEE; 2020. p. 1624–8.
21. Mustaqeem A, Anwar SM, Majid M. Multiclass classification of cardiac arrhythmia using improved feature selection and SVM invariants. *Comput Math Methods Med.* 2018;2018. Article ID 7310496.
22. Mustaqeem A, Anwar SM, Majid M. A modular cluster based collaborative recommender system for cardiac patients. *Artif Intell Med.* 2020;102:101761.
23. Bodén AC, Molin J, Garvin S, West RA, Lundström C, Treanor D. The human-in-the-loop: an evaluation of pathologists’ interaction with AI in clinical practice. *Histopathology.* 2021;79:210.
24. Goyal D. Medical image segmentation using interactive refinement (Doctoral dissertation, Arizona State University). 2021.
25. Holzinger A, Malle B, Saranti A, Pfeifer B. Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI. *Inf Fusion.* 2021;71:28–37.
26. Pearl J. *Causality.* Cambridge University Press; 2009.

-
27. Garreau D, Luxburg U. Explaining the explainer: a first theoretical analysis of LIME. In: International conference on artificial intelligence and statistics. PMLR; 2020. p. 1287–96.
 28. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Computer vision – ECCV 2014. Cham: Springer International Publishing; 2014. p. 818–33.
 29. Chen J, Song L, Wainwright M, Jordan M. Learning to explain: an information-theoretic perspective on model interpretation. In: International conference on machine learning. PMLR; 2018. p. 883–92.
 30. Aas K, Jullum M, Løland A. Explaining individual predictions when features are dependent: more accurate approximations to Shapley values. arXiv preprint arXiv:1903.10464, 2019.
 31. Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825, 2017.
 32. Montavon G, Binder A, Lapuschkin S, Samek W, Müller K-R. Layer-wise relevance propagation: an overview. In: Explainable AI: interpreting, explaining and visualizing deep learning. Springer; 2019. p. 193–209.
 33. Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One. 2015;10(7):e0130140.
 34. Samek W, Montavon G, Lapuschkin S, Anders CJ, Müller K-R. Explaining deep neural networks and beyond: a review of methods and applications. Proc IEEE. 2021;109(3):247–78.



Optimizing Radiologic Detection of COVID-19

36

Test Set Technologies and Artificial Intelligence

Z. Gandomkar, P. C. Brennan, and M. E. Suleiman

Contents

Introduction	511
Test Set Technologies	512
Artificial Intelligence in Medical Imaging and the Opportunity Provided by Test Sets	512
Why Does Education Need to Embrace AI to Enhance Detection of COVID-19?	514
AI-Tailored Education Using Clinician Demographics and Image Features .. .	514
AI Companions to Help with COVID-19 Diagnosis	516
Finally, Some Cautionary Notes	517
Conclusion	517
References	517

Abstract

COVID-19 has had a huge impact globally. This chapter examines the role that test set technologies coupled with artificial intelligence can transform clinical educational strategies. Using AI to streamline a methodology that has been around for decades enables education that is tailored to each clinician, acknowledges each individual's weaknesses,

and is available instantly wherever in the world the clinician is available. Cautionary notes are also provided.

Keywords

Test set technologies · COVID-19 · AI,
Diagnostic performance

Introduction

At the time of writing, COVID-19 has infected 62 million individuals, respectively, according to official reports [1]. The World Health Organization has stated that up to 35% of infected individuals can have long-term symptoms

Z. Gandomkar · P. C. Brennan (✉) · M. E. Suleiman
University of Sydney, Sydney, NSW, Australia
e-mail: ziba.gandomkar@sydney.edu.au;
patrick.brennan@sydney.edu.au;
moe.suleiman@sydney.edu.au

(3 weeks post-testing) with 20% of 18- to 34-year-olds reporting prolonged symptoms [2]. The former amounts to up over 20 million individuals, which exceeds the incidence rates of all cancers combined [3]. In addition, it is important to emphasize that the US Centers for Disease Control and Prevention refer to the fact that there may be between 6x and 24x more actual cases than reported numbers, with the actual number fluctuating widely across locations [4].

High-resolution lung-computed tomography (CT) is the frontline tool facilitating accurate assessment of disease severity, progression, or treatment response [5–7]. However, recent research done by our group has shown that clinicians' sensitivity using HRCT for COVID-19 can be 50% or lower (discussed more below): We need better image interpretations [8]. Educational solutions with and without artificial intelligence (AI) to optimize COVID-19 diagnostic competency are required. One of the key approaches focuses on using test set technologies that are revolutionizing diagnoses in other domains such as breast cancer, lung cancer, and dust diseases, blending in an unusual way the latest in artificial intelligence, human expertise, and ergonomics.

Test Set Technologies

What is a test set technology? It is often a web-based program allowing clinicians and trainees ("readers") to diagnose sets of radiologic images wherever in the world they are located [9]. These clinically relevant test cases, which in the current context would be CT images of the chest to diagnose COVID-19, allows readers to look at real de-identified medical images the same way as they would in the clinic. A reader independently judges each image in a test set and whenever relevant identifies and locates COVID-19 lesions such as ground-glass opacities, mosaic patterns, or consolidations. The reader then decides using a confidence scale as to how sure they are that a COVID-19 infection exists. Readers can go back to any image or case and correct a previous decision prior to submitting their answers. Some cases contain disease while

others are COVID-19 free (Fig. 1). Instantly following such analyses, the most modern systems intelligently analyze the data and present a range of performance values including sensitivity and specificity to each clinician. In addition, using novel AI algorithms, reader-specific image files will be generated so that correct and incorrect decisions can be examined and education test sets can be built and delivered that are most appropriate to each individual radiologist. The result is a better way for radiologists to effectively improve their cancer detection skills.

There is good evidence that such intelligent educational platforms work. As reported in the world's leading journals such as Academic Radiology, this type of approach improves mean levels of disease detection by 34%, with individual clinicians achieving higher rates [10, 11]. This dramatic improvement occurs regardless of expertise and experience. In addition, these self-assessment approaches are popular and in terms of adoption, 90% of relevant clinicians in specific jurisdictions have been shown to voluntarily engage. For example, within 6 weeks of the launch of the CovED tool by the University of Sydney start-up company – DetectED-X, there were users in 145 countries. Partnering with major agencies such as GE and Amazon, to ensure the widest distribution, the highest levels of security and rapid access are critically important.

Artificial Intelligence in Medical Imaging and the Opportunity Provided by Test Sets

Artificial intelligence (AI), machine learning (ML), and deep learning (DL) are terms that are often used interchangeably. However, AI is a general term that can be interpreted as integrating human intelligence to machines, while ML is a subset of AI, which focuses on giving computers the ability to learn from a dataset (such as images within a test set), and DL is a technique used to accomplish ML. Nonetheless, to realize powerful AI systems that can accurately perform inferences from a large set of data, just as humans, AI systems need to learn from accurate and large enough

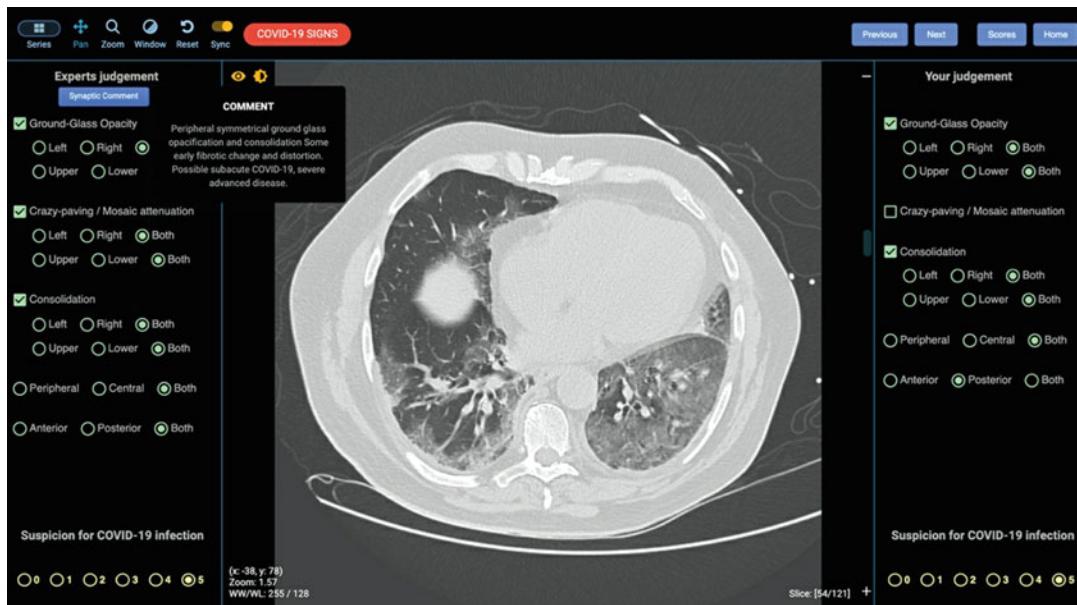


Fig. 1 An example of a test set image presented *after* the clinician has judged the image. On the right-hand side are the radiologists' classification of appearances and their location, on the left-hand side are the expert radiologists'

classification of appearances and their comments on the case. The disease appearances in this case were correctly identified. These images are from the recently developed CovED tool

training datasets. AI models are fed experiences through vast amounts of information with the aim of producing correct predictions for new unseen data, hence the training datasets need to be accessible, useable, valuable, and regularly validated.

Medical imaging providers generate huge amounts of data containing valuable information; such data however is often left in storage systems unutilized due to the regulatory restrictions around ethics and privacy. Furthermore, data extracted from the clinical world usually inherits the messiness of the real world reducing its usability and value. Nonetheless, increased usability and value can be accomplished by annotating and labeling the data to create large numbers of training samples; however, this would introduce yet another obstacle as expert annotators are expensive and mostly not available.

The test set technology introduced above provides radiologists with a more accessible and organized opportunity to utilize AI to improve their performance in detecting abnormalities in medical images. To be able to provide accurate test sets, medical images have to be anonymized,

annotated, and labeled with a summary of radiologists' reports; furthermore, this information is validated through further testing and diagnosis. Hence, if we take breast cancer diagnosis, for example, test sets containing cases on cancer will curate vast amounts of highly accurate information and data regarding the lesions' exact location (biopsy proven), appearance on the mammogram, and further information on past mammograms; this type of information along with the radiologists specifically recorded interactions with the images can help train AI systems to detect areas of interest on mammograms more accurately, giving radiologists much needed help on confirming a diagnosis.

Another example of the test set offering is around COVID-19 diagnosis. Although the appearances of COVID-19 are now known, they overlap with other pneumonitis diagnosis, making an accurate diagnosis harder. Through accurately diagnosed test sets, supported by high-quality pathology data, large number of cases with confirmed diagnosis of COVID-19 and accurate information on the locations of the main appearances of

the disease can be made available. This is like gold dust when attempting to train AI systems on the diagnosis of COVID-19. Furthermore, radiologists' interactions with the test set are also extensively recorded, which creates information on the difficulty of certain appearances, the presence of other diseases, and the effect of different technologies on the performance of radiologists on the diagnosis of COVID-19. This will help create AI systems that would customize educational tools to concentrate on radiologists' weaknesses, for example, allowing them to be faster and more confident in diagnosing COVID-19.

The remaining part of this chapter will further explain the need for such intelligent education and the methods used to ensure that artificial intelligence is the core technology to optimizing educational strategies when diagnosing COVID-19.

Why Does Education Need to Embrace AI to Enhance Detection of COVID-19?

Industry and researchers have looked at the extent of clinician efficacy when diagnosing COVID-19 from CT scans and a paper produced at the end of 2020 and currently in submission highlights the issue [8]. In brief the work showed for each of the three key COVID-19 CT features, many images were given a normal (no disease) rating, even though according to the experts, that appearance was there.

Two key findings were:

- Sensitivities as low as 65%, 25%, and 35% for ground-glass opacities, consolidation, and crazy paving appearances, respectively.
- Current levels of experience, education, fellowship expertise, specialization, and familiarity with COVID-19 had no impact on performance.

The conclusion from that work was that despite current methods of formal or informal education, clinicians are failing to recognize key radiologic features associated with COVID-19. Innovative educational solutions are required. In particular, it must be acknowledged that if we are to use test

set approaches as described above to enhance COVID-19 education, it must be acknowledged that educational modules must be tailored to suit the specific weakness of each user. It is not enough to simply make available the same modules for everyone since Dr. Brown's errors will be different from Dr. Li's. In this chapter, therefore, we describe a novel artificial intelligence solution to tailor the educational experience for each clinician or health care worker. The method allows specific algorithms for each individual to be built as they interact with the CovED images, resulting in solutions that can be available within a matter of minutes.

AI-Tailored Education Using Clinician Demographics and Image Features

Error making patterns of radiologists have been widely studied in the medical image perception sciences. It was shown that individual radiologist's error making patterns are consistent across cases [12, 13]. Therefore, a machine could be able to successfully learn clinicians' weakness and strengths and tailor educational materials for each individual. In the era of precision medicine, such strategy for precision medical education is highly desirable. Figure 2 shows possible approaches for using AI in order to customize the educational materials.

Considering the nature of the pandemic and clinicians' lack of time to spend on learning about the CT appearances of COVID-19, tailored educational platform is even more important than before. Previously, in other imaging modalities several attempts have been made by researchers to predict radiologists' errors to retrieve difficult cases for radiologists. As an example, in [14], an AI model relies on the previous interactions of the radiologist with mammograms and predicts if a new case is difficult for that radiologist. The model relies on a set of computer-extracted features from mammograms and trains two classifiers based on these features to classify the input images as false positives or true negatives (normal cases) as well false negative or true positives (cancer cases) [14]. The group extended these AI

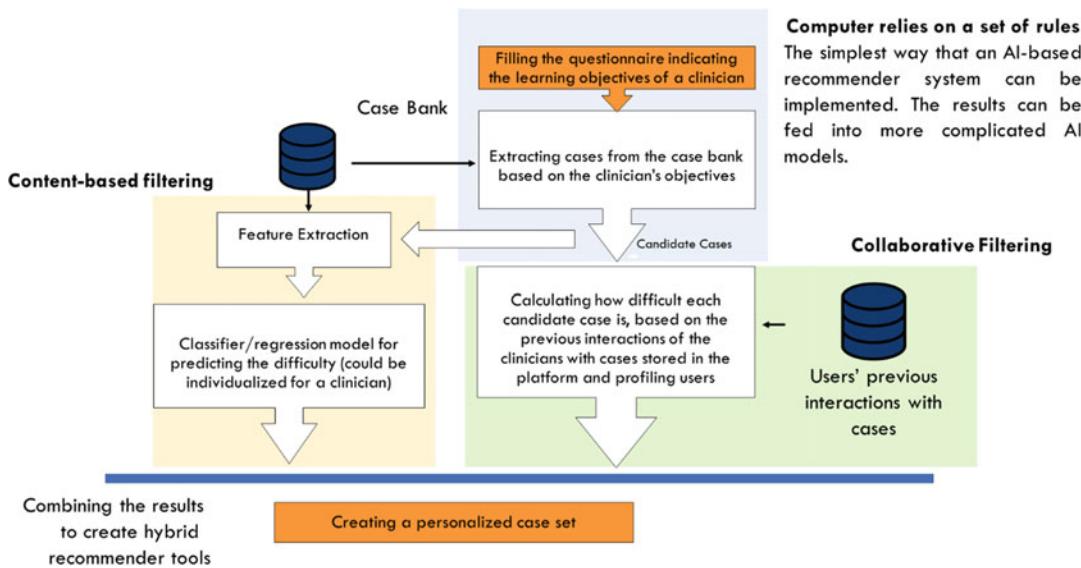


Fig. 2 Possible ways of leveraging the artificial intelligence (AI) algorithms for creating tailored educational materials

models to include other features in [15]. Another approach could be modeling the problem as a recommender system, which predicts user's rating to a case based on ratings of similar users to the same case or similar cases.

Collecting high-quality data is at the core of establishing such AI tools for personalized education. This requires platforms for recording radiologists' interactions with cases, so that the error-making patterns of observes can be modeled. As an example, using the large amount of data collected by the CovED platform [16], such an AI model has been built. In the platform, clinicians were asked to provide a score of 0 to 5, showing their confidence about the presence of COVID-19. The platform recorded all radiologists' interactions with the system. A solution, which predicts difficult positive and negative cases specific to each clinician, was developed based on a matrix factorization-based recommender system [17]. In summary, the clinician case rating matrix (\tilde{P}), where each row represented a radiologist and each column represented a case, was generated.

The clinician case rating matrix was decomposed into the product of two lower dimensionality matrices, i.e., $\tilde{P} = RC$, where R and C represent radiologist and case profiles in a new

space of latent factors. Intuitively, through the user and case profiles, the recommender system models similarities between cases and users. The predicted rating, $p_{i,j}^i$, from radiologist i to case j can be formulated as

$$p_j^i = \sum_{f=0}^{nfactors} R_{if} C_{fj}$$

where f is an iterator on the latent factors while $nfactors$ shows the number of factors. The number of latent factors is a hyper-parameter of our model, which should be set beforehand. It was shown that a matrix factorization with one latent factor is equivalent to the average score to the items without any personalization [17]. To improve the personalization, the number of latent factors should be increased. However, by increasing number of latent factors, models start to overfit to the training data and hence the quality of its prediction to a blinded test data will drop [17]. Therefore, a nice balance to avoid overfitting while reaching proper level of personalization is desirable. Usually, $nfactors$ should be set empirically. In these recommender systems, handling data sparsity is important as for each clinician assessments to a number of cases will be missing.

To handle the data sparsity, various methods have been proposed in the literature [17].

The matrix-factorization-based recommender system described above using the CovED platform achieved an AUC of 0.80 (CI:0.78–81) and 0.83 (CI:0.82–0.84) in detecting difficult positive and negative cases, respectively. Further analysis revealed that when ratings from 15 clinicians were available for a case, the mean absolute error for predicting the rating was, on average, ≤ 1 . Also, availability of a clinician's ratings for 15 cases guaranteed an average mean absolute error of approximately ≤ 1 in predicting the ratings of that clinician for new cases.

Therefore, adequate number of readings should be available for each user and each case and collecting interaction data is at the center of establishing such a model. Although the proposed methodology has been developed for COVID-19 cases, it can be extended to personalize other radiologic learning materials.

AI Companions to Help with COVID-19 Diagnosis

Various AI tools have been developed since beginning of the pandemic to aid clinicians to diagnose COVID-19 on chest CTs [18], segment the extent of infections [19], and classify COVID-19 severity [20]. The majority of these tools rely on deep learning models, whose main bottleneck is availability of high-quality datasets. Many clinicians from around the world attempted to make publicly available COVID-19 datasets, such as:

1 – “MosMedData: Chest CT Scans with COVID-19 Related Findings [21].” It includes 254 normal cases and various CT manifestation of COVID-19: ground-glass opacifications, where involvement of lung parenchyma is less than 25% (684 cases); ground-glass opacifications where involvement of lung parenchyma is between 25 and 50% (125 cases); ground-glass opacifications and regions of consolidation, where involvement of lung parenchyma is between 50 and 75% (45 cases); diffuse ground-glass opacifications and consolidation as well as reticular changes in lungs, where involvement of lung parenchyma exceeds 75% (2 cases). It also includes 50 cases,

where masks corresponding to the ground-glass opacifications and consolidations is provided.

2 – “China National Center for Bioinformation [22].” It includes 750 CT slices from 150 COVID-19 patients were manually segmented into background, lung field, ground-glass opacity, and consolidation.

3 – “UCSD dataset (confirmed by Tongji Hospital radiologist) [23].” It includes 350 CT slices from 216 COVID-19 patients with captions describing CT features and their extents. It also provides 396 CT slices from 170 non-COVID-19 patients.

4 – “COVID-19 CT segmentation dataset (<https://medicalsegmentation.com/covid19/>).”

It includes three sets of CT images with masks for ground-glass opacities, consolidations, and pleural effusion for 473 slices of 19 patients.

5 – “COVID-19 CT Lung and Infection Segmentation Dataset [24].” It includes lung and infection masks for 20 CT cases (left lung, right lung, and infections are labeled by two radiologists and verified by an experienced radiologist).

6 – “BIMCV COVID-19+ [25].” It includes 163 CT cases with image-level labels. It also provides 5 CT studies with semantic segmentation of radiographic findings.

Although many AI studies aimed at classifying images as COVID-19 or non-COVID-19 images, due to the similarity between COVID-19 CT manifestation and other viral pneumonia, radiology professional bodies recommend against using CT as the primary diagnostic tool [26]. The reverse transcriptase polymerase chain reaction (RT-PCR) assay of nasal and pharyngeal swab specimens is the gold standard for identifying SARS-CoV-2 patients [26]. However, lung CT images have been used for clinical triage of symptomatic patients who underwent chest CT following emergency room to speed up diagnostic workflow and establish isolation at admission. More importantly, assessing CT manifestation of COVID-19 independently predicted an adverse patient outcome associated with COVID-19 pneumonia [27]. It also can help for clinical staging of the disease and patients with milder symptoms and comorbidities [28]. Considering potential application of the CT to managing the COVID-19, deep learning-based segmented infectious areas as a

mask or heatmap could be the most useful “companions” for radiologists in COVID-19 pandemic.

Various tools for segmenting the areas infected by COVID-19 in CT have been proposed. As an example, Inf-Net [29] is a state-of-the-art deep learning model for segmenting the infection areas. Another example is COPLE-Net [30], which proposes a noise-robust dice loss function to address the challenges introduced by the noise in annotations. Deep learning models were also fine-tuned for multiclass segmentation to segment ground-glass opacities, consolidation, and pleural effusion.

Considering the paucity of fully annotated COVID-19 CT images, semi-supervised learning methods can also help in improving the performance of the deep learning models by leveraging availability of unannotated data [29, 31]. Usually in a semi-supervised approach, unannotated COVID-19 CT slices can be used to produce pseudo-annotated data [31]. To avoid overfitting, this can be done iteratively as follows:

- Inputs: Trained networks for segmenting COVID-19 on CT slices and unannotated COVID-19 data
- Repeat until no unannotated image is left
- Randomly select N CT images, annotate them using the network (pseudo-label set)
- Retrain the network using the union of original annotated images and this new set
- Omit the randomly selected images from the set containing unannotated data

For the images in the pseudo-annotation set, soft labels can be used. Soft labels provide a membership probability to a point rather than a binary number shows if the pixel is a member of the class or not. In the datasets, where a text is available for describing the CT features, the value of the soft label will be adjusted based on the image-level-provided labels.

Finally, Some Cautionary Notes

It should be noted that certain unintended consequences could occur as a result of introducing an AI companion tool to radiologists. For example, the perception of a “safety-net” could lead to a less

thorough visual search pattern as a result of over-trusting the AIs [32], an effect possibly more prominent among less-experienced radiologists. Also, introducing an AI might result in an extensive level of fatigue as readers might end up searching twice (or more), thus exhibiting prolonged fixations or multiple refixations on locations previously ruled out by their peripheral vision and potentially with a large proportion of these being normal aberrations [33]. The need therefore for clinicians to be educated thoroughly about the proper way to interact with AIs is required to establish and maintain the appropriate level of trust between the human and machine. Using heatmaps, for example, to explain AI decisions might facilitate such trust.

Conclusion

We have hit an important milestone regarding clinical education. Using the COVID-19 paradigm and novel AI algorithms, researchers and pedagogists can now develop systems of education that are tailored to the individual, rapidly accelerating learning activities. While the focus here is COVID-19 and radiology, the lessons learned can be applied across educational domains.

References

1. WHO Coronavirus Disease (COVID-19) Dashboard. 30th November 2020.
2. Tenford MW, et al. Symptom duration and risk factors for delayed return to usual health among outpatients with COVID-19 in a multistate health care systems network – United States, March–June 2020. Morb Mortal Wkly Rep (MMWR). 2020;69:993–8.
3. Cancer Research UK. Worldwide cancer statistics. 2018. [online] Available at: <https://www.cancerresearchuk.org/health-professional/cancer-statistics/worldwide-cancer#:~:text=There%20were%2017%20million%20new,of%20all%20cancers%20diagnosed%20worldwide>. [Accessed 15 December 2020].
4. Havers FP, Reed C, Lim T, Montgomery JM, Klena JD, Hall AJ, Fry AM, Cannon DL, Chiang CF, Gibbons A, Krapivunaya I. Seroprevalence of antibodies to SARS-CoV-2 in 10 sites in the United States, March 23–May 12, 2020. JAMA Intern Med. 2020;180:1776.
5. Revel MP, Parkar AP, Prosch H, Silva M, Sverzellati N, Gleeson F, Brady A. European Society of Radiology (ESR) and the European Society of Thoracic Imaging

- (ESTI). COVID-19 patients and the radiology department – advice from the European Society of Radiology (ESR) and the European Society of Thoracic Imaging (ESTI). *Eur Radiol.* 2020;30(9):4903–4909. <https://doi.org/10.1007/s00330-020-06865-y>. Epub 2020 Apr 20. PMID: 32314058; PMCID: PMC7170031.
6. Liu KC, Xu P, Lv WF, et al. CT manifestations of coronavirus disease-2019: a retrospective analysis of 73 cases by disease severity. *Eur J Radiol.* 2020;126:108941. <https://doi.org/10.1016/j.ejrad.2020.108941>.
 7. Francone M, Iafrate F, Masci GM, Coco S, Cilia F, Manganaro L, Panebianco V, Andreoli C, Colaiacomo MC, Zingaropoli MA, Ciardi MR, Mastroianni CM, Pugliese F, Alessandri F, Turriziani O, Ricci P, Catalano C. Chest CT score in COVID-19 patients: correlation with disease severity and short-term prognosis. *Eur Radiol.* 2020;30(12):6808–6817. <https://doi.org/10.1007/s00330-020-07033-y>. Epub 2020 Jul 4. PMID: 32623505; PMCID: PMC7334627.
 8. Gandomkar Z, Suleiman M, Brennan PC, et al. Even experts might fail in recognising CT manifestation of COVID-19. Submitted for publication 2020.
 9. Brennan PC, Lee W, Tapia K. Breast Screen Reader Assessment Strategy (BREAST): a research infrastructure with a translational objective. In: Samei E, Krupinski E, editors. *The handbook of medical image perception and techniques*. 2nd ed. Cambridge University Press; 2019.
 10. Suleiman W, Rawashdeh M, Lewis S, McEntee M, Lee W, Tapia K, Brennan P. Impact of breast reader assessment strategy on mammographic radiologists' test reading performance. *J Med Imaging Radiat Oncol.* 2016;60(3):352–8.
 11. Trieu Y, Tapia KB, Frazer H, Lee W, Brennan PC. Improvement of cancer detection on mammograms via BREAST test sets. *Acad Radiol.* 2019;26:e341–7.
 12. Gandomkar Z, Tay K, Brennan PC, Mello-Thoms C. Recurrence quantification analysis of radiologists' scanpaths when interpreting mammograms. *Med Phys.* 2018;45(7):3052–62.
 13. Gandomkar Z, Tay K, Ryder W, Brennan PC, Mello-Thoms C. iCAP: an individualized model combining gaze parameters and image-based features to predict radiologists' decisions while reading mammograms. *IEEE Trans Med Imaging.* 2016;36(5):1066–75.
 14. Mazurowski MA, Baker JA, Barnhart HX, Tourassi GD. Individualized computer-aided education in mammography based on user modeling: concept and preliminary experiments. *Med Phys.* 2010;37(3):1152–60.
 15. Grimm LJ, Ghate SV, Yoon SC, Kuzniak CM, Kim C, Mazurowski MA. Predicting error in detecting mammographic masses among radiology trainees using statistical models based on BI-RADS features. *Med Phys.* 2014;41(3):031909.
 16. Suleiman ME, Rickard M, Brennan PC. Perfecting detection through education. *Radiography.* 2020;26: S49–53.
 17. Falk K. Practical recommender systems. Manning Publications; 2019.
 18. Ozsahin I, Sekeroglu B, Musa MS, Mustapha MT, Uzun OD. Review on diagnosis of COVID-19 from chest CT images using artificial intelligence. *Comput Math Methods Med.* 2020;2020:9756518.
 19. Zhang HT, Zhang JS, Zhang HH, Nan YD, Zhao Y, Fu EQ, Xie YH, Liu W, Li WP, Zhang HJ, Jiang H. Automated detection and quantification of COVID-19 pneumonia: CT imaging analysis by a deep learning-based software. *Eur J Nucl Med Mol Imaging.* 2020;47(11):2525–32.
 20. Zhu J, Shen B, Abbasi A, Hoshmand-Kochi M, Li H, Duong TQ. Deep transfer learning artificial intelligence accurately stages COVID-19 lung disease severity on portable chest radiographs. *PLoS One.* 2020;15(7):e0236621.
 21. Morozov S, Andreychenko A, Pavlov N, Vladzymyrskyy A, Ledikhova N, Gombolevskiy V, et al. MosMedData: chest CT scans with COVID-19 related findings dataset. 2020.
 22. Zhang K, Liu X, Shen J, et al. Clinically Applicable AI System for Accurate Diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography [published correction appears in *Cell.* 2020;182(5):1360]. *Cell.* 2020;181(6):1423–1433.e11. <https://doi.org/10.1016/j.cell.2020.04.045>.
 23. Zhao J, Zhang Y, He X, Xie P. Covid-ct-dataset: a ct scan dataset about covid-19. 2020. Preprint at <https://arxiv.org/abs/2003.13865>.
 24. Ma J, Wang Y, An X, Ge C, Yu Z, Chen J, et al. Towards efficient COVID-19 CT annotation: a benchmark for lung and infection segmentation in arXiv:2004.12537. 2020. [online] Available: <http://arxiv.org/abs/2004.12537>.
 25. Vayá MdII, Saborit JM, Montell JA, Pertusa A, Bustos A, Cazorla M, Galant J, Barber X, Orozco-Beltrán D, García F, et al. Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients. arXiv preprint arXiv:2006.01174. 2020.
 26. Raptis CA, Hammer MM, Short RG, Shah A, Bhalla S, Bierhals AJ, Filev PD, Hope MD, Jeudy J, Kligerman SJ, Henry TS. Chest CT and coronavirus disease (COVID-19): a critical review of the literature to date. *Am J Roentgenol.* 2020;215(4):839–842. <https://doi.org/10.2214/AJR.20.23202>. Epub 2020 Apr 16. PMID: 32298149.
 27. Meiler S, Schaible J, Poschenrieder F, Scharf G, Zeman F, Rennert J, Pregler B, Kleine H, Stroszczynski C, Zorger N, Hamer OW. Can CT performed in the early disease phase predict outcome of patients with COVID 19 pneumonia? Analysis of a cohort of 64 patients from Germany. *Eur J Radiol.* 2020;131:109256.
 28. Feng Z, Yu Q, Yao S, Luo L, Zhou W, Mao X, Li J, Duan J, Yan Z, Yang M, Tan H. Early prediction of disease progression in COVID-19 pneumonia patients with chest CT and clinical characteristics. *Nat Commun.* 2020;11(1):1–9.
 29. Fan D-P, Zhou T, Ji G-P, Zhou Y, Chen G, Fu H, et al. Inf-net: automatic COVID-19 lung infection segmentation from CT images. 2020.
 30. Wang G, Liu X, Li C, Xu Z, Ruan J, Zhu H, Meng T, Li K, Huang N, Zhang S. A noise-robust framework for

- automatic segmentation of COVID-19 pneumonia lesions from CT images. *IEEE Trans Med Imaging*. 2020;39(8):2653–2663. <https://doi.org/10.1109/TMI.2020.3000314>. PMID: 32730215.
31. Li Y, Chen J, Xie X, Ma K, Zheng Y. Self-loop uncertainty: a novel pseudo-label for semi-supervised medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Champions: Springer; 2020. p. 614–23.
32. Du-Crow E, Warren L, Astley SM, Hulleman J. Is there a safety-net effect with computer-aided detection (CAD)? In: Medical imaging 2019: image perception, observer performance, and technology assessment, vol. 10952. International Society for Optics and Photonics; 2019. p. 109520J.
33. Taylor-Phillips S, Stinton C. Fatigue in radiology: a fertile area for future research. *Br J Radiol*. 2019;92(1099):20190043.



Clare McGenity, Alex Wright, and Darren Treanor

Contents

Introduction	522
Definitions	523
A Brief History	523
Clinical Applications	524
Detecting and Classifying Disease	525
Grading and Scoring Disease	525
Finding or Outlining Tumors and Tissue	525
Finding Rare Events and Small Objects	527
Predictive Tasks	527
Image Quality Tools	528
Grand Challenges	528
Technical Aspects of Digital Pathology	528

Supplementary Information: The online version of this chapter (https://doi.org/10.1007/978-3-030-64573-1_278) contains supplementary material, which is available to authorized users.

C. McGenity (✉) · A. Wright
Leeds Teaching Hospitals NHS Trust, Leeds, UK

University of Leeds, Leeds, UK
e-mail: clare.mcgenity@nhs.net; A.Wright@leeds.ac.uk

D. Treanor
Leeds Teaching Hospitals NHS Trust, Leeds, UK

University of Leeds, Leeds, UK
Department of Clinical Pathology and Department of
Clinical and Experimental Medicine, Linköping
University, Linköping, Sweden

Centre for Medical Image Science and Visualization
(CMIV), Linköping University, Linköping, Sweden
e-mail: darrentreanor@nhs.net

Standards of Reporting of AI	531
Deployment and Regulation	532
Technical	532
Infrastructure	532
People	533
Education	533
Conclusion	533
Cross-References	533
References	534

Abstract

Artificial intelligence is already impacting many areas of society, and there is much opportunity for future change and benefit to our lives. Surgical pathology is no exception to this technological shift. The growing availability of digital pathology is facilitating the development of algorithms to tackle challenging or laborious aspects of the pathologist's assessment and diagnosis. However, there are many issues to overcome before we see widespread, routine use of AI in clinical practice. In this chapter, we explore a brief history, clinical applications, technical aspects, standards of reporting, deployment, regulation, and education opportunities of AI in surgical pathology. We provide the reader with a general overview of the present position of surgical pathology in the world of AI.

Keywords

Surgical pathology · Histopathology · Pathology · Artificial intelligence · Machine learning · Deep learning · Digital pathology · Whole slide imaging · Virtual pathology · Image analysis

Introduction

Artificial intelligence (AI) and specifically deep learning (DL) are emerging technologies in surgical pathology, with much scope to impact clinical practice. There are 367,167 new cases of cancer per year in the UK [1], and

approximately 45% of cancer patients have surgery to remove their tumor [2]. Every one of these patient's biopsies or surgical specimens is examined and reported by pathologists. In addition to these figures, there are many more diagnostic and screening tests for cancer and specimens for nonmalignant diseases that are processed by pathology departments routinely. As such, a revolution in this field would significantly impact both patients and healthcare systems. While the concept of using computers to analyze images is not new to surgical pathology, the increasing availability of digital pathology scanning has created the ideal landscape for the development of AI.

In this chapter, we explore a brief history, clinical uses, technical considerations, standards of reporting, deployment, regulation, and educational potential of AI in surgical pathology. This chapter is intended for a scientific audience not familiar with AI, and its purpose is to be an overview of developments in this rapidly growing field, rather than an exhaustive review of all the literature.

Development and deployment of AI on a large scale has become a realistic possibility in recent years due to increasing computing power, capacity to store large images, and the expanding use of whole slide imaging to create digital images [3]. Whole slide imaging (Figs. 1 and 2) involves scanning of an entire glass slide and storing the image electronically at high resolution so that it is available for review on a computer [3]. The availability of these high-resolution medical images has created a platform for AI innovation.

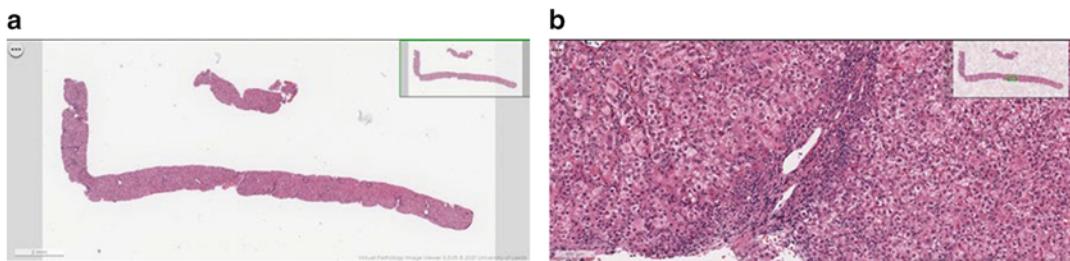


Fig. 1 (a) Whole slide image of liver biopsy at low power. (b) Whole slide image of liver biopsy at high power [4]

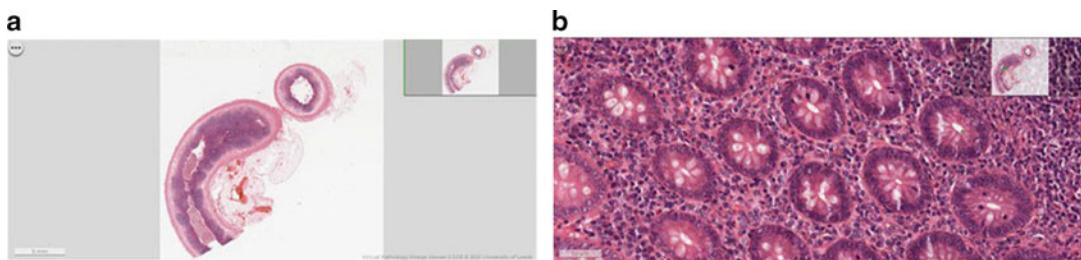


Fig. 2 (a) Whole slide image of appendix specimen at low power. (b) Whole slide image of appendix specimen at high power [4]

Definitions

- Artificial intelligence – the development of computers able to engage in humanlike thought processes [5].
- Deep learning – a set of methods using the combination of multiple nonlinear transformations, aiming to create abstract, and potentially useful, representations. It is a branch of machine learning [6].
- Digital pathology a generic term for the use of electronic images in pathology, usually facilitated by whole slide imaging technology [7].
- Image analysis – process of inputting an image to an operator, and a measurement is produced as the output. Examples include automated detection and diagnosis of disease, identifying a specified region, size measurements, and calculation of risk [8].
- Machine learning – an area of artificial intelligence that involved creating applications that learn and improve their accuracy over time without specific programming that makes them perform this process [9].

- Telepathology – the projection of surgical pathology images electronically from one location to another for diagnosis and assessment by a pathologist [7].
- Virtual microscopy – the conversion of histological sections on glass microscope slides to high resolution digital images [10].
- Whole slide imaging – a technology that is used to digitize a glass slide in its entirety, making the slide available electronically for assessment by a pathologist as a digital image [7].

A Brief History

Artificial intelligence has created much excitement and interest in recent years, but the concept of intelligent machines is not a new one. Notably, Alan Turing published his groundbreaking paper “Computing Machinery and Intelligence” over 70 years ago, in 1950 [11], and such concepts even had their place in ancient mythology with humanlike, intelligent machines fashioned by the Greek god Hephaestus [12].

The origins of AI in pathology can be dated back as far as the mid-seventeenth century to Antoni van Leeuwenhoek who successfully created a system for measuring microscopic objects such as red blood cells, considered to be the early beginnings of image analysis [13]. However, it was not until there was significant advancement in computing technology in the mid-twentieth century that the path to modern digital pathology and AI truly started [14]. An early development in this period was the CELLSCAN system presented at a computing conference in 1961 that described the use of a television microscope and computer to perform analysis of white blood cells [15]. Another example from this era was the computer-aided image analysis of cells in a blood smear by Prewitt and Mendelsohn in 1966 [16]. They used the combination of a flying spot microscope, digital converter, and magnetic tape recorder to scan and record cell features. The magnetic tape was read by a computer to produce a grey scale image of the cell, reflecting cell borders and simple characteristics.

Telepathology was both the precursor to the development of digital pathology and remains a pathologist's tool in its own right. Ronald Weinstein, a pioneer of telepathology, wrote an editorial on its creation and uses in 1986 [17]. The technology allows assessment of tissue without the pathologist needing to be physically present with a glass microscope slide. It has two primary uses: first for frozen sections of a patients' tissue, requiring a diagnosis or decision to be made intraoperatively, and second to facilitate an expert second opinion on a case [18]. The first virtual microscope was developed in 1996 by the Saltz

group who applied technology to histology slides that was originally intended for photomicrographs [18, 19]. The first patents for whole slide imaging were filed in 1997 and 1998 by Bacus Research Laboratory [20]. Initially, the digitization of glass slides was limited by slow scanning capability, the need for focus stacking, and poor interoperability of the equipment for digitization [21]. Whole slide imaging has since evolved and overcome these issues with much improved scanning technology (Fig. 2) and can now rapidly produce high quality images, thus making the deployment of digital pathology more practical [20, 22]. The utility of deep learning algorithms in other fields, combined with the growing availability of digital pathology, has formed a natural progression toward applying AI to tasks in surgical pathology [3] (Fig. 3).

Clinical Applications

AI is predicted to spread across many healthcare settings; however, research in this area remains at the early stages [24]. In surgical pathology, universal clinical adoption of AI algorithms is reliant on routine digitization of histology slides, and this process has happened comparatively slowly when contrasted with digitization in other medical specialties such as radiology [24]. Despite both enthusiasm for this technology and concern that it may replace pathologists, current developments are narrow and task based and do not yet reflect the complexity of assessment and diagnosis by pathologists in the real world [25]. A statement by Prof. Geoffrey Hinton in 2016 is worth noting, "People should stop training radiologists now. It's



Fig. 3 Example set up of multiple whole slide imaging scanners [23]

just completely obvious that within 5 years, deep learning is going to do better than radiologists,” and yet, there is still a definite need for training radiologists at this time [26]. Recent data in the UK shows a declining workforce with only 3% of surgical pathology departments having enough staff to meet current clinical demand [27], and therefore, task-based AI could actually be used to assist busy pathologists in a range of ways. Presently, research focusses on more common cancers such as prostate cancer, breast cancer, and lung cancer, although interest in less common cancers and nonneoplastic disease is growing [28, 29]. We explore examples of algorithms applied to different task groups for digital pathology. We have divided these tasks into detecting and classifying disease, grading and scoring disease, finding rare events and small objects, and finding or outlining tumors, predictive tasks, and image quality improvement. Additionally, we will review the recent rise in grand challenges to solve diagnostic problems using AI in pathology.

Detecting and Classifying Disease

The pathologist’s role is important in the diagnosis of many diseases, but the process is not always simple. There are often multiple components, and this is only expanding with the evolving role of molecular pathology [30, 31]. An area targeted by researchers is the detection and classification of disease, and there are examples across multiple pathological specialties. Table 1 summarizes examples of papers reporting models that performed detection or classification tasks. Breast pathology is a subspecialty where deep learning algorithms have been successfully developed for such purposes in breast cancer [32–34]. For example, one algorithm was able to sort cases into nonmalignant or malignant with an AUC of 0.962 and three class accuracy of 81.3% for classification into normal/benign, DCIS, and invasive ductal carcinoma groups [33]. Similarly, AI has been applied to examples in uropathology, with research focusing primarily on prostate cancer [35–37]. One interesting study gathered a very

large dataset with a total of 44,732 whole slide images (WSIs) to perform classification and detection of prostate cancer, and also skin and breast cancers [35]. This group achieved an area under the curve (AUC) of 0.98 for the three cancer types. These same concepts can be applied to nonmalignant disease, and an example is a model developed for duodenal biopsies that distinguished between coeliac disease, nonspecific duodenitis, and normal tissue with >0.95 AUC for all cases [38]. Deployment of such tools could assist with prioritizing the most urgent or severe cases and potentially achieve faster turnaround times for those that are most critical.

Grading and Scoring Disease

The process of providing a diagnosis may involve providing a grade or score for aspects of a disease. This is usually an indication of severity and can impact the clinical management of the patient. Grading and scoring tasks can be laborious and susceptible to variability between pathologists [44, 45], and so this an area where AI may assist. Algorithms for Gleason grading in prostate cancer achieved a good performance in assessing prostate biopsies [46, 47]. One model with a dataset of over 1200 digital slides achieved a quadratic kappa score of 0.918 for Gleason grading of prostate cancer biopsies, indicating very good agreement. Furthermore, this concept may be applied to scoring of immunohistochemistry, and examples of this can be seen in breast pathology [48, 49]. One group achieved an overall accuracy of 83% when scoring HER2 immunohistochemistry in breast cancer [49]. Examples of models performing grading and scoring tasks are summarized in Table 2.

Finding or Outlining Tumors and Tissue

It may be useful to identify areas of malignant tumor in digital slides to highlight these features to the pathologist, and therefore assist in their diagnosis. Some examples of algorithms performing these tasks are summarized in Table 3.

Table 1 Examples of models performing detection and classification of disease

Reference	Disease/condition	Task	Performance measure	Dataset numbers	Staining
Araujo et al. [32]	Breast cancer	Classification of breast cancer	Carcinoma/noncarcinoma accuracy 83.3% and four class accuracy 77.8%	269 WSIs	H&E
Bejnordi et al. [33]	Breast cancer	Classification of breast cancer	AUC of 0.962 for classification of nonmalignant and malignant and three-class accuracy of 81.3%	221 WSIs	H&E
Bulten et al. [37]	Prostate cancer	Classification into malignant and nonmalignant	F1 score 0.62 discriminating tumor and nontumor	94 WSI pairs (H&E and IHC)	H&E, IHC
Campanella et al. [35]	Prostate cancer, basal cell carcinoma, and breast cancer metastases	Diagnosis of multiple cancers	AUC above 0.98 for all cancer types	44,732 WSIs: 24,859 prostate images, 9,962 skin images, and 9,894 breast images	H&E
Cruz-Roa et al. [34]	Breast cancer	Detection of invasive breast cancer	Dice coefficient of 75.86%, PPV 71.62%, and NPV 96.77% pixel-by-pixel evaluation compared to manually annotated invasive ductal carcinoma	605 cases	H&E
Halicek et al. [39]	Squamous cell carcinoma (SCC) of head and neck and thyroid cancer	Head and neck cancer detection	AUC of 0.916 for SCC group and AUC 0.954 for thyroid carcinoma group for detection of tumor	381 WSIs	H&E
Litjens et al. [36]	Prostate cancer and breast cancer sentinel nodes	Prostate cancer detection and metastatic breast cancer detection	AUC (median analysis) 0.99 for prostate cancer and 0.88 for sentinel node metastases	225 prostate WSIs +173 breast WSIs	H&E
Noorbaksh et al. [40]	Multiple cancers	Classification of 19 solid tumor types into malignant and benign	AUC 0.995(+/- 0.008)	27,815 WSIs	H&E
Sharma et al. [41]	Gastric cancer	Detecting cancer and detecting necrosis	Classification accuracy of 0.6990 for cancer and 0.8144 for necrosis	454 cases	H&E, IHC
Wei et al. [42]	Lung cancer	Classification of lung adenocarcinoma subtypes	AUC greater or equal to 0.97 for all classes	422 WSIs	H&E
Wei et al. [38]	Celiac disease	Classification of duodenal biopsies	Accuracies of 95.3% for coeliac disease, 91% for normal tissue, and 89.2% for nonspecific duodenitis. AUC>0.95 for all cases	1230 WSIs	H&E
Zhang et al. [43]	Bladder cancer	Bladder cancer diagnosis	AUC 97% and mean accuracy 94.6%	913 WSIs	H&E

Table 2 Examples of models grading and scoring disease

Reference	Disease/condition	Task	Performance measure	Dataset numbers	Staining
Bulten et al. [46]	Prostate cancer	Gleason grading of prostate cancer	Deep learning system agreement with reference standard – quadratic Cohen's kappa of 0.918, AUC 0.990 for benign versus malignant, AUC 0.978 grade group 2 or more, and 0.974 grade group 3 or more	1243 WSIs	H&E
Ertosun et al. [50]	Brain cancer	Glioma grading	Classification accuracy of 71% for identifying the grade of LGG into Grade II or Grade III	54 WSIs	H&E
Qaiser et al. [48]	Breast cancer	Immunohistochemical scoring of HER2 status	Mean scoring accuracy for four classes ranged 0.65–0.794	172 WSIs	H&E, IHC
Ström et al. [47]	Prostate cancer	Gleason grading of prostate cancer	Average pairwise kappa for assigning Gleason grades of 0.62	6953 WSIs for training. Test set of 1631 biopsies and external validation set of 330 biopsies	H&E
Vandenbergh et al. [49]	Breast cancer	Immunohistochemical scoring of HER2 status	Overall accuracy 83% between AI and pathologist	71 cases	IHC

Table 3 Examples of models finding or outlining tumors

Reference	Disease/condition	Task	Performance measure	Dataset numbers	Staining
Argarwalla et al. [51]	Breast cancer	Tumor segmentation	Best reported F1 score 0.83	270 WSIs	H&E
Liang et al. [52]	Gastric cancer	Tumor segmentation	Mean accuracy 91.09%	1900 images	H&E
Jia et al. [53]	Colon cancer	Tumor segmentation	Best reported F-measures of 0.836 and 0.835 for cancer images	910 images	H&E
Qaiser et al. [54]	Colon cancer	Tumor segmentation	Accurate tumor segmentation on two datasets – F1 scores 0.9243 and 0.8273, respectively	125 WSIs	H&E

Finding Rare Events and Small Objects

Finding and highlighting small objects or rare events may assist the pathologist with a diagnosis of a range of diseases. Table 4 summarizes examples of models that identified these features.

Predictive Tasks

A compelling use for some AI applications is the ability to perform tasks that may be challenging,

subjective, or impossible for humans. AI may be manipulated to make a range of predictions from digital slides, including predicting grade, subtype, survival, disease recurrence, and mutation status. One model was able to successfully predict cases of biochemical recurrence of prostate cancer within 1 year, performing with an AUC of 0.845 on external validation [59]. In another example, an algorithm used for breast cancer cases was able to predict tumor grade with 82% accuracy, ER status with 84% accuracy, and 75% for a risk of recurrence score [60]. Using AI to make

Table 4 Examples of models finding rare events or small objects

Reference	Disease/condition	Task	Performance measure	Dataset numbers	Staining
Kashif et al. [55]	Colon cancer	Cell detection	F1 score 0.748	15 images	H&E
Tellez et al. [56]	Breast cancer	Mitosis detection	F1score 0.480	832 WSIs	H&E, PHH3
Wang et al. [57]	Lung cancer	Cell detection	F1 score 0.8215	300 lung cancer tiles	H&E
Xing et al. [58]	Brain tumor, pancreatic cancer, and breast cancer	Nuclei segmentation	Mean F1 score for nuclei detection: 0.77 brain tumor 0.88 pancreatic NET 0.78 breast cancer	31 brain tumor images 22 pancreatic NET images 35 breast cancer images	H&E

predictions, especially the molecular profile of a disease from H&E slides, could in future save time and money that are spent on additional tests, in turn, potentially reaching a faster diagnosis and progressing the patient's management at an earlier stage. The model capable of predicting ER status is one illustration of this, and a further example is the paper that described successfully predicting six lung cancer mutations with H&E slides [61]. This was performed with an AUC of 0.733–0.856 for mutations STK11, EGFR, FAT1, SETBP1, KRAS, and TP53, therefore potentially reducing the need for additional testing of these in the future. Table 5 summarizes predictive artificial intelligence tools.

Image Quality Tools

AI has been applied to other aspects of the surgical pathology workflow. A convolutional neural network was constructed to identify and quantify out of focus areas created during the slide digitization process and could be used to trigger a rescan of the slide before review by a pathologist if needed [65]. A further example explored image quality and the impact of image artifacts on machine learning, for a clinical trial dataset, and found a minimal increase in accuracy when these cases were removed [66]. Image quality is a growing area of interest for the application of artificial intelligence.

Grand Challenges

One approach to speeding up the development of digital pathology AI algorithms has been to create publicly available datasets and to challenge international researchers to develop solutions [67]. Examples include the Camelyon dataset which has been successfully used to create AI tools capable of detecting metastatic breast cancer in lymph nodes (Fig. 4) [68]. Another challenge focused on distinguishing between glioblastoma multiforme and low-grade glioma brain tumors [69]. A recent challenge called the Prostate cANcer graDe Assessment (PANDA) challenge was in progress at the time of writing and consists of a dataset of 11,000 WSIs of H&E-stained prostate biopsies [70]. There is a growing number of similar challenges with publicly available datasets across multiple medical specialties [71] (Fig. 5).

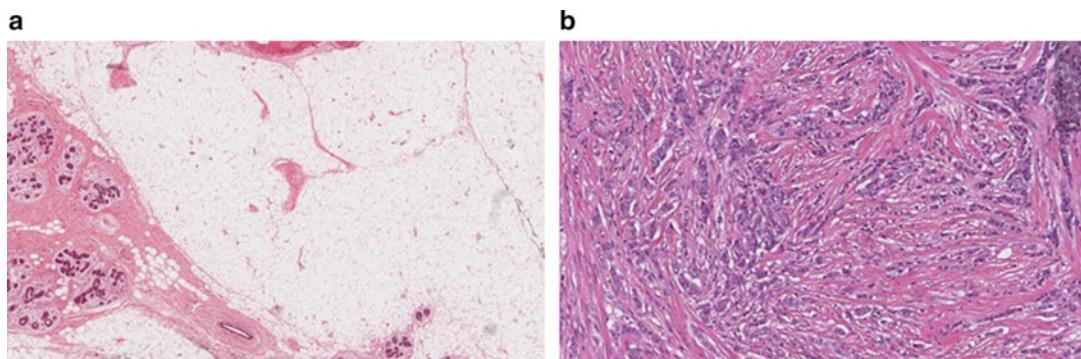
Technical Aspects of Digital Pathology

Digital pathology is the process of creating digital slides, and it involves four key stages:

1. Capturing the image
2. Storing the image
3. Manipulation and editing
4. Displaying and viewing the image [72]

Table 5 Examples of models performing predictive tasks

Reference	Disease/condition	Task	Performance measure	Dataset numbers	Staining
Coudray et al. [61]	Lung cancer	Genomics prediction from pathology images	AUC range 0.733–0.856 for six mutations	1634 WSIs +340 slides	H&E
Couture et al. [60]	Breast cancer	Prediction of cancer grade, ER status, subtypes, and recurrence risk	Accuracies were 82% for predicting grade, 84% for ER status, 77% for basal-like versus nonbasal like, 94% for ductal versus lobular, and 75% for high versus low-medium ROR-PT score	1203 cases	H&E, IHC
Kather et al. [62]	Colon cancer	Survival prediction of colorectal cancer	Demonstrated deep stroma score significantly prognostic of survival in all tumor stages HR 1.99 (1.27–3.12) whereas pathologist annotations were not in any stage	1382 WSIs	H&E
Tang et al. [63]	Brain cancer and lung cancer	Prediction of survival time of </= 1 year or >1 year	Best reported AUC for glioblastoma was 0.722 Best reported AUC for lung cancer was 0.702	424 WSIs of glioblastoma 305 WSIs of lung cancer	H&E
Veta et al. [64]	Breast cancer	Predict mitotic score and PAM50 proliferation score	Best reported mitotic score prediction: quadratic-weighted Cohen's kappa score $\kappa = 0.567$, 95% CI [0.464, 0.671] predicted scores versus ground truth Best proliferation score prediction: Spearman's correlation coefficient of $r = 0.617$, 95% CI [0.581 0.651] predictions versus ground truth	821 WSIs	H&E
Yamamoto et al. [59]	Prostate cancer	Predict prostate cancer recurrence	AUC 0.845 for predicting biochemical recurrence within 1 year on external validation	15,464 WSIs	H&E

**Fig. 4** (a) Digital slide of normal, benign breast tissue. (b) Digital slide of invasive breast cancer [4]

The availability of whole slide imaging technology that can rapidly perform these steps has made routine use of this a possibility in pathology laboratories [22].

Whole slide images (WSIs) are captured using digital slide scanners, typically using 20x or 40x objective lens which creates images of approximately 0.5 and 0.25 microns per pixel (MPP),

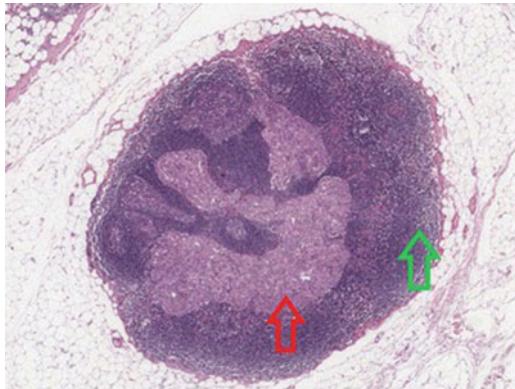


Fig. 5 Lymph node showing metastatic breast cancer (red arrow) and background normal lymph node tissue (green arrow) [4]

respectively [73, 74]. WSIs can vary significantly in size depending on the tissue scanned, the magnification used, and the image compression type [75]. As an example, a typical digital slide scanned at 20x (0.5 MPP) is approximately one gigapixel in size, and using JPEG2000 compression at image quality 70 requires between 200 and 400 megabytes of storage. Therefore, digital pathology imaging is highly demanding on image storage and archiving systems, and routine scanning can generate prohibitively large quantities of data, unless accounted for prior to deployment [76]. Currently, there is no common standard digital slide format [77], with each digital slide scanner vendor producing their own files, and often requiring their own software to view and analyze the images. However, collaborative work is being undertaken to extend the DICOM file format to incorporate digital slide images and pathology data [78]. Until this standard is achieved, many third party software solutions exist that can be used to view and analyze digital slides from multiple vendors [79].

The digitization of glass slides also introduces the possibility of image analysis and automated quantitation [77]. WSIs are stored as matrices containing red green blue (RGB) values describing the color of each pixel in the image, with three 8-bit values per pixel. It is these numeric values that form the basis of all subsequent automation and means that one digital slide can contain billions of values. The size of WSIs requires that

analysis must be done either at a lower power (zoomed out) or in parts – referred to as tiles, patches, or blocks – and aggregated in the final stages of processing [80]. Traditionally, image analysis in digital pathology used “low-level” techniques such as pixel-level thresholding and binary object morphology, to create hand-crafted image features that were fed into AI algorithms such as support vector machines (SVMs) and random forests, for predicting tissues classifications [80]. These methods have been rapidly phased out of digital pathology research due to the rising popularity of deep learning [29, 36].

Deep learning is a form of machine learning, which is in turn a subtype of AI, frequently used in digital pathology [28]. Deep learning is a form of artificial neural network (ANN) that uses auto-encoders to circumvent the traditional ANN issue known as the vanishing gradient problem [81]. This allows networks to be built using more layers (“deeper”), allowing more granular levels of representation, and subsequently higher prediction accuracy. Convolutional neural networks (CNNs) provide an additional framework prior to the ANN, whereby the pixel data from images are distilled into high level representations using three different processes: convolution, activation, and pooling [82]. The aim of these processes is to create an algorithm capable of interpreting raw input data (pixels in this case) by changing to more abstract representations of the data over multiple layers to make it more interpretable and identify important features [6, 83]. It uses a technique called back-propagation between layers of data to learn from looking at its past function and adjusts its internal parameters to minimize the error rate of each function’s prediction [84]. There are several phases to developing a deep learning algorithm [85] (Fig. 6):

- Data collection and formatting – deep learning algorithms require many training examples to ensure they are generalizable to real-world data [86]. Typically, networks are trained using labeled examples of images (supervised learning), and all images that pass through the network need a uniform size [87].
- Network architecture – the structure of the network should be designed according to the

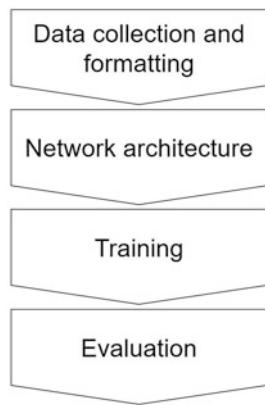


Fig. 6 Flowchart of key phases in developing a deep learning algorithm

visual task. This includes the arrangement of network layers, number of nodes per layer, and the convolution, activation, and pooling operations (if using a CNN). Alternatively, transfer learning is a popular method that takes existing network structures and retrains the algorithm on new training data [88].

- Training – the algorithm passes labeled examples through the network and adjusts its parameters based on their ability to correctly predict the image classification. Once every single image in the training set has passed through the network, the algorithm has processed one epoch. The number of epochs and other settings such as learning rate and batch size are known as hyperparameters, and these need to be optimized by the algorithm developer in order to maximize performance [89].
- Evaluation – deep learning algorithm design typically incorporates model evaluation or testing, which requires part of the dataset to be withheld from the training process [90]. The resulting model accuracy is then calculated from its predictive ability on the data that is not used for training. Cross validation is a popular evaluation strategy that iterates through the dataset in folds, so that each image is used in the testing process at least once [91].

Deep learning has seen a rapid rise in popularity due to the relative ease at which algorithms can be developed and deployed, as well as

outperforming traditional AI methodologies [92]. However, when deploying algorithms on real-world data, performance can be adversely affected by many factors:

- Overfitting – if the algorithm is not exposed to a training dataset that represents all the possible variations of images that it is likely to encounter, it will underperform on images that it has not seen before [93]. Algorithms need to be trained on a comprehensive dataset that has been carefully curated prior to processing.
- Variability in digital slides – image quality can affect algorithm performance by creating unnecessary variation to the appearance of digital slides [94]. Variation can occur as a digital or histological artifact and has the capacity to affect AI significantly [66]. Standardization of glass slide production and robust QC processes during scanning can help to mitigate this issue [95, 96], and color normalization is a popular tool for correcting stain variation after scanning [97].
- Human error in gold standard – due to the large volumes required, training examples that are hand-labeled inevitably contain human error [98]. These errors have less impact on algorithm performance in larger training datasets, but where it is infeasible to generate large numbers of training examples, it is considered good practice to double-score images.

Standards of Reporting of AI

A search of clinicaltrials.gov in December 2020 for terms “artificial intelligence,” “deep learning,” and “machine learning” revealed over 1000 registered clinical trials. In this growing research field, concern about the reporting and evaluation of AI technologies in healthcare is ongoing. To tackle this, Parikh et al. [99] proposed five criteria to consider in the evaluation of the quality of predictive algorithms in 2019. These were as follows:

1. Meaningful endpoints (e.g., patient survival, positive predictive value, or sensitivity)

2. Appropriate benchmarks (e.g., against “a clinician’s best judgment”)
3. Interoperable and generalizable (otherwise, these may be barriers to use by a clinician)
4. Specify interventions (e.g., the outcome of the algorithm links to an intervention to improve patient care)
5. Audit mechanisms (algorithms should be subject to ongoing audit after regulatory approval)

Following this, the provision of more extensive checklists was proposed to reflect the complexity and range of considerations for these technologies. The CONSORT-AI and SPIRIT-AI guidelines were released in 2020 as extensions of the existing Equator Network guidelines for clinical trials [100, 101]. These adaptations were developed in recognition of the differences in AI compared to other interventions, and as an attempt to standardize reporting of AI and improve transparency. The CONSORT-AI and SPIRIT-AI extensions contain 14 and 15 checklist items, respectively. These include the following:

- Stating the intended use of the AI intervention
- Giving the clinical context in which it should be used
- Details of the data inputs and expected outputs
- Details of the human-AI interaction
- How the outputs will impact decision-making or other aspects of clinical care
- How errors are handled

Additionally, further Equator network guideline extensions for AI were under development at the time of writing. The STARD-AI and TRIPOD-AI will be extensions of the current guidance for studies of diagnostic accuracy and multivariate prediction models, respectively [102, 103].

Deployment and Regulation

The widespread adoption of artificial intelligence in pathology is yet to happen, but a fully digitized workflow is generally considered a requirement to achieve this [104]. There are examples of many

centers who have already deployed digital pathology [105, 106], and there are recognized benefits to digitization in addition to its use for AI [76]. There are multiple considerations in the successful clinical adoption of AI. These may be broadly grouped into technical, infrastructure, or people-related issues, and some of these are listed below:

Technical

- To produce useful outputs, good quality input data is required, and even small changes in the input data when an algorithm is deployed can result in a reduction in performance [107].
- Current algorithms for surgical pathology are task orientated and rely on large amounts of training data. While there are no agreed optimal dataset sizes for training AI in digital pathology, the consensus is that to create robust algorithms that are transferable to routine pathology image data, algorithms should be trained on enough data to model such variation [67].

Infrastructure

- It is already difficult to manage large volumes of image data in digital pathology, but even more so additionally managing large numbers of extracted image features. This is a recognized challenge, along with having the systems able to cope with the data, algorithms, and their interaction [20].
- Digital pathology and AI may be expensive for laboratories to implement together at a good standard with systems that are adequately fast [67].
- AI tools for pathology will likely be introduced across multiple competing platforms, which will create issues around interoperability of the technology [108].
- Having the facilities for adequate visualization and visual analytics is vital in successfully interpreting imaging features of AI in digital pathology [20].

People

- There is much enthusiasm for AI in surgical pathology, but these systems are difficult to successfully deploy, and some may fail. It is important to maintain engagement from stakeholders when challenges are faced during the implementation of a new technology [67].
- Pathologists must have enough confidence in algorithms to sign out reports while using them, and there are new legal considerations associated with this [108].
- Involvement of members of the multi-disciplinary team in the development of AI could help in the practical implantation of algorithms into a pressured healthcare system [109].

Regulated clinical adoption of AI in healthcare is an area of growing importance, with an urgent need for a robust regulatory framework for tools that do not currently fit the measures in place [110]. Regulatory bodies are recognizing this need and starting to adapt to these technologies [111]. Despite the challenges, the first AI algorithms have received regulatory approval for use in clinical practice in pathology. Tools for supporting pathologists to detect prostate cancer [112] and PD-L1 positivity in lung cancer [113] are examples of algorithms with European CE-IVD approval. Similarly, there are examples with the US Food and Drug Administration (FDA) clearance for a breast cancer diagnosis algorithm [114] and algorithms for the diagnosis of multiple cancer subtypes [115]. Further guidance is likely to appear from regulatory bodies as more of these technologies are developed.

Education

Digital pathology and AI have created new ways of presenting surgical pathology education to medical students, specialty trainees, and for training and conferences in general [3, 116]. A practical use of the accumulation of the large

collections of cases for training and validating AI is that these could also be accessed for educational purposes [116]. The COVID-19 pandemic has highlighted the value of digital pathology use for training as it is very amenable to use remotely via secure videoconferencing software [117]. Some pathologists have created automated annotations and other interactive functions for trainees with this technology [3, 109]. Automated diagnostic aids could also support independent reporting of cases by surgical pathology trainees and biomedical scientists [109]. Finally, a general area for consideration is using education to build trust with pathologists, which could be achieved by adapting current training so that it provides a basic understanding of how these tools function and how to incorporate them into practice [108].

Conclusion

The landscape of AI in surgical pathology is changing continually, with new advancements appearing all the time. There has been significant progress in this field; however, there are still many barriers to overcome before AI becomes an established component of pathological diagnosis. Nevertheless, it is an exciting and dynamic area, and we are likely to see a transformation in the use of this technology in surgical pathology.

Cross-References

- ▶ [AIM in Genomics](#)
- ▶ [Artificial Intelligence in Kidney Pathology](#)

Acknowledgments Dr. McGenity is funded by Leeds Hospitals Charity <https://www.leedshospitalscharity.org.uk/>. Dr. Wright and Dr. Treanor are funded by National Pathology Imaging Co-operative (NPIC). NPIC (project no. 104687) is supported by a £50 m investment from the Data to Early Diagnosis and Precision Medicine strand of the Government's Industrial Strategy Challenge Fund, managed and delivered by UK Research and Innovation (UKRI).

References

1. Cancer Research UK: Cancer statistics for the UK. <https://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk#heading-Four> (2020). Accessed 4 Jan 2021.
2. Cancer Research UK: Cancer treatment statistics. <https://www.cancerresearchuk.org/health-professional/cancer-statistics/treatment#heading-One> (2020). Accessed 4 Jan 2021.
3. Niazi MKK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. *Lancet Oncol.* 2019;20(5):e253–e61. [https://doi.org/10.1016/S1470-2045\(19\)30154-8](https://doi.org/10.1016/S1470-2045(19)30154-8).
4. Division of Pathology & Data Analytics, Leeds Institute of Medical Research, University of Leeds: Virtual Pathology Slide Library. <https://www.virtualpathology.leeds.ac.uk/slides/library/index.php> (2021). Accessed 18 Jan 2021.
5. Kok JN, Boers E, Kosters WA, Van der Putten P, Poel MJ. Artificial intelligence: definition, trends, techniques, and cases. *Artif Intell.* 2009;1:1–20.
6. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell.* 2013;35(8):1798–828.
7. Cross S, Furness P, Igali L, Snead D, Treanor D. Best practice recommendations for implementing digital pathology. London: The Royal College of Pathologists; 2018. p. 3–5.
8. Kallergi M. Evaluation strategies for medical-image analysis and processing methodologies. In: Costaridou L, editor. *Medical image analysis methods.* 1st ed. Florida: CRC Press; 2005. p. 434.
9. IBM: Machine Learning. <https://www.ibm.com/cloud/learn/machine-learning> (2020). Accessed 13 Jan 2021.
10. Mikula S, Trots I, Stone JM, Jones EG. Internet-enabled high-resolution brain mapping and virtual microscopy. *NeuroImage.* 2007;35(1):9–15. <https://doi.org/10.1016/j.neuroimage.2006.11.053>.
11. Turing AM. Computing machinery and intelligence. *Mind.* 1950;LIX(236):433–60. <https://doi.org/10.1093/mind/LIX.236.433>.
12. Nelson R. From Hephaestus' automatons to OpenAI's deep learning. *EE-Evaluation Engineering;* 2016. p. 2.
13. Meijer GA, Beliën JA, van Diest PJ, Baak JP. Origins of ... image analysis in clinical pathology. *J Clin Pathol.* 1997;50(5):365–70. <https://doi.org/10.1136/jcp.50.5.365>.
14. Aeffner F, Zarella MD, Buchbinder N, Bui MM, Goodman MR, Hartman DJ, et al. Introduction to digital image analysis in whole-slide imaging: a white paper from the digital pathology association. *J Pathol Inform.* 2019;10:9.
15. Preston K. The CELLSCAN system – T.M. a leucocyte pattern analyzer. Papers presented at the May 9–11, 1961, western joint IRE-AIEE-ACM computer conference. Los Angeles: Association for Computing Machinery; 1961. p. 173–83.
16. Prewitt JMS, Mendelsohn ML. The analysis of cell images*. *Ann N Y Acad Sci.* 1966;128(3):1035–53. <https://doi.org/10.1111/j.1749-6632.1965.tb11715.x>.
17. Weinstein RS. Prospects for telepathology. *Hum Pathol.* 1986;17(5):433–4. [https://doi.org/10.1016/s0046-8177\(86\)80028-4](https://doi.org/10.1016/s0046-8177(86)80028-4).
18. Kayser K, Kayser G, Radziszowski D, Oehmann A. From telepathology to virtual pathology institution: the new world of digital pathology. *Romanian J Morphol Embryol.* 1999;45:3–9.
19. Ferreira R, Moon B, Humphries J, Sussman A, Saltz J, Miller R, et al. The virtual microscope [conference paper]. In: Proc AMIA Annu Fall Symp. 1997 October 25–29, Nashville, Tennessee. JAMIA, symposium supplement, 449–453.
20. Pantanowitz L, Sharma A, Carter A, Kurc T, Sussman A, Saltz J. Twenty years of digital pathology: an overview of the road travelled, what is on the horizon, and the emergence of vendor-neutral archives. *J Pathol Inform.* 2018;9(1):40. https://doi.org/10.4103/jpi.jpi_69_18.
21. Weinstein RS, Holcomb MJ, Krupinski EA. Invention and early history of telepathology (1985–2000). *J Pathol Inform.* 2019;10:1. https://doi.org/10.4103/jpi.jpi_71_18.
22. Pantanowitz L, Valenstein P, Evans A, Kaplan K, Pfeifer J, Wilbur D, et al. Review of the current state of whole slide imaging in pathology. *J Pathol Inform.* 2011;2(1):36. <https://doi.org/10.4103/2153-3539.83746>.
23. Division of Pathology & Data Analytics, Leeds Institute of Medical Research, University of Leeds: The Virtual Pathology Research Section at Leeds. <https://www.virtualpathology.leeds.ac.uk/research/> (2021). Accessed 18 Jan 2021.
24. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25(1):44–56. <https://doi.org/10.1038/s41591-018-0300-7>.
25. Cheng JY, Abel JT, Balis UGJ, McClinton DS, Pantanowitz L. Challenges in the development, Deployment & Regulation of Artificial Intelligence (AI) in Anatomical Pathology. *Am J Pathol.* 2020. <https://doi.org/10.1016/j.ajpath.2020.10.018>.
26. European Society of Radiology. What the radiologist should know about artificial intelligence – an ESR white paper. *Insights Imag.* 2019;10(1):44.
27. The Royal College of Pathologists: The Pathology Workforce. <https://www.rcpath.org/discover-pathology/public-affairs/the-pathology-workforce.html> (2020). Accessed 8 Dec 2020.
28. Acs B, Rantalaainen M, Hartman J. Artificial intelligence as the next step towards precision pathology. *J Intern Med.* 2020;288(1):62–81. <https://doi.org/10.1111/joim.13030>.
29. Srinidhi CL, Ciga O, Martel AL. Deep neural network models for computational histopathology: a survey. *J Med Image Anal.* 2020;67:101813.
30. Saikia B, Gupta K, Saikia UN. The modern histopathologist: in the changing face of time. *Diagn Pathol.*

- 2008;3(1):25. <https://doi.org/10.1186/1746-1596-3-25>.
31. Moch H, Blank PR, Dietel M, Elmberger G, Kerr KM, Palacios J, et al. Personalized cancer medicine and the future of pathology. *Virchows Arch.* 2012;460(1):3–8. <https://doi.org/10.1007/s00428-011-1179-6>.
32. Araújo T, Aresta G, Castro E, Rouco J, Aguiar P, Eloy C, et al. Classification of breast cancer histology images using Convolutional Neural Networks. *PLoS One.* 2017;12(6):e0177544. <https://doi.org/10.1371/journal.pone.0177544>.
33. Bejnordi BE, Zuidhof G, Balkenhol M, Hermsen M, Bult P, van Ginneken B, et al. Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images. *J Med Imag.* 2017;4(4):044504.
34. Cruz-Roa A, Gilmore H, Basavanhally A, Feldman M, Ganesan S, Shih NNC, et al. Accurate and reproducible invasive breast cancer detection in whole-slide images: a deep learning approach for quantifying tumor extent. *Sci Rep.* 2017;7(1):46450. <https://doi.org/10.1038/srep46450>.
35. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med.* 2019;25(8):1301–9. <https://doi.org/10.1038/s41591-019-0508-1>.
36. Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci Rep.* 2016;6(1):26286. <https://doi.org/10.1038/srep26286>.
37. Bulten W, Litjens G. Unsupervised prostate cancer detection on H&E using convolutional adversarial autoencoders. *arXiv* 2018(preprint arXiv:07098).
38. Wei JW, Wei JW, Jackson CR, Ren B, Suriawinata AA, Hassanpour S. Automated detection of celiac disease on duodenal biopsy slides: a deep learning approach. *J Pathol Inform.* 2019;10:7.
39. Halicek M, Shahedi M, Little JV, Chen AY, Myers LL, Sumer BD, et al. Head and neck cancer detection in digitized whole-slide histology using convolutional neural networks. *J Sci Rep.* 2019;9(1):1–11.
40. Noorbakhsh J, Farahmand S, Soltanieh-ha M, Namburi S, Zarringhalam K, Chuang J. Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images. *Nat Commun* 2020;11:6367. <https://doi.org/10.1038/s41467-020-20030-5>.
41. Sharma H, Zerbe N, Klempert I, Hellwisch O, Hufnagl P. Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology. *Comput Med Imaging Graph.* 2017;61:2–13.
42. Wei JW, Tafe LJ, Linnik YA, Vaickus LJ, Tomita N, Hassanpour SJ Sr. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Sci Rep.* 2019;9(1):1–8.
43. Zhang Z, Chen P, McGough M, Xing F, Wang C, Bui M, et al. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nat Mach Intell.* 2019;1(5):236–45.
44. de Vet HCW, Knipschild PG, Schouten HJA, Koudstaal J, Kwee W-S, Willebrand D, et al. Interobserver variation in histopathological grading of cervical dysplasia. *J Clin Epidemiol.* 1990;43(12):1395–8. [https://doi.org/10.1016/0895-4356\(90\)90107-Z](https://doi.org/10.1016/0895-4356(90)90107-Z).
45. Ozkan TA, Eruyar AT, Cebeci OO, Memik O, Ozcan L, Kuskonmaz I. Interobserver variability in Gleason histological grading of prostate cancer. *Scand J Urol.* 2016;50(6):420–4. <https://doi.org/10.1080/21681805.2016.1206619>.
46. Bulten W, Pinckaers H, van Boven H, Vink R, de Bel T, van Ginneken B, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol.* 2020;21(2):233–41.
47. Ström P, Kartasalo K, Olsson H, Solorzano L, Delahunt B, Berney DM, et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *Lancet Oncol.* 2020;21(2):222–32.
48. Qaiser T, Rajpoot NM. Learning where to see: a novel attention model for automated immunohistochemical scoring. *IEEE Trans Med Imaging.* 2019;38(11):2620–31.
49. Vandenberghe ME, Scott ML, Scorer PW, Söderberg M, Balcerzak D, Barker C. Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer. *Sci Rep.* 2017;7(1):1–11.
50. Ertosun MG, Rubin DL. Automated grading of gliomas using deep learning in digital pathology images: a modular approach with ensemble of convolutional neural networks. *AMIA Annual Symposium Proceedings.* American Medical Informatics Association; 2015. p. 1899.
51. Agarwalla A, Shaban M, Rajpoot NM. Representation-aggregation networks for segmentation of multi-gigapixel histology images. *arXiv* 2017(preprint arXiv:08814).
52. Liang Q, Nan Y, Coppola G, Zou K, Sun W, Zhang D, et al. Weakly supervised biomedical image segmentation by reiterative learning. *IEEE J Biomed Health Inform.* 2018;23(3):1205–14.
53. Jia Z, Huang X, Eric I, Chang C, Xu Y. Constrained deep weak supervision for histopathology image segmentation. *IEEE Trans Med Imaging.* 2017;36(11):2376–88.
54. Qaiser T, Tsang Y-W, Taniyama D, Sakamoto N, Nakane K, Epstein D, et al. Fast and accurate tumor segmentation of histology images using persistent homology and deep convolutional features. *Med Image Anal.* 2019;55:1–14.
55. Kashif MN, Raza SEA, Sirinukunwattana K, Arif M, Rajpoot N. Handcrafted features with convolutional neural networks for detection of tumor cells in histology images. 2016 IEEE 13th International

- Symposium on Biomedical Imaging (ISBI). IEEE; 2016. p. 1029–32.
56. Tellez D, Balkenhol M, Otte-Höller I, van de Loo R, Vogels R, Bult P, et al. Whole-slide mitosis detection in H&E breast histology using PHH3 as a reference to train distilled stain-invariant convolutional networks. *IEEE Trans Med Imaging*. 2018;37(9):2126–36.
 57. Wang S, Yao J, Xu Z, Huang J. Subtype cell detection with an accelerated deep convolution neural network. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2016. p. 640–648.
 58. Xing F, Xie Y, Yang L. An automatic learning-based framework for robust nucleus segmentation. *IEEE Trans Med Imaging*. 2015;35(2):550–66.
 59. Yamamoto Y, Tsuzuki T, Akatsuka J, Ueki M, Morikawa H, Numata Y, et al. Automated acquisition of explainable knowledge from unannotated histopathology images. *Nat Commun*. 2019;10(1):1–9.
 60. Couture HD, Williams LA, Geraerts J, Nyante SJ, Butler EN, Marron J, et al. Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. *NPJ Breast Cancer*. 2018;4(1):1–8.
 61. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med*. 2018;24(10):1559–67.
 62. Kather JN, Krisam J, Charoentong P, Luedde T, Herpel E, Weis C-A, et al. Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. *PLoS Med*. 2019;16(1):e1002730.
 63. Tang B, Li A, Li B, Wang M. CapSurv: capsule network for survival analysis with whole slide pathological images. *IEEE Access*. 2019;7:26022–30.
 64. Veta M, Heng YJ, Stathonikos N, Bejnordi BE, Beca F, Wollmann T, et al. Predicting breast tumor proliferation from whole-slide images: the TUPAC16 challenge. *Med Image Anal*. 2019;54:111–21.
 65. Kohlberger T, Liu Y, Moran M, Chen P-H, Brown T, Hipp J, et al. Whole-slide image focus quality: automatic assessment and impact on ai cancer detection. *J Pathol Inform*. 2019;10(1):39. https://doi.org/10.4103/jpi.jpi_11_19.
 66. Wright AI, Dunn CM, Hale M, Hutchins G, Treanor D. The effect of quality control on accuracy of digital pathology image analysis. *IEEE J Biomed Health Inform*. 2020;1. <https://doi.org/10.1109/JBHI.2020.3046094>.
 67. Tizhoosh HR, Pantanowitz L. Artificial intelligence and digital pathology: challenges and opportunities. *J Pathol Inform*. 2018;9:38.
 68. Bejnordi BE, Veta M, Van Diest PJ, Van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*. 2017;318(22):2199–210.
 69. Xu Y, Jia Z, Wang L-B, Ai Y, Zhang F, Lai M, et al. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. *BMC Bioinform*. 2017;18(1):1–17.
 70. Wouter Bulten GL, Pinckaers H, Ström P, Eklund M, Egevad L, Grönberg H, Kartasalo K, Ruusuvuori P, Häkinen T, Dane S, Demkin M. Prostate cANcer graDe Assessment (PANDA) Challenge. <https://panda.grand-challenge.org/> (2020). Accessed 5 Jan 2021.
 71. Ginneken BV, Kerkstra S, Meakin J. Challenges. <https://grand-challenge.org/challenges/> (2020). Accessed 9 Dec 2020.
 72. Pantanowitz L. Digital images and the future of digital pathology. *J Pathol Inform*. 2010;1:15. <https://doi.org/10.4103/2153-3539.68332>.
 73. Treanor D. Virtual slides: an introduction. *Diagn Histopathol*. 2009;15(2):99–103.
 74. Sellaro TL, Filkins R, Hoffman C, Fine JL, Ho J, Parwani AV, et al. Relationship between magnification and resolution in digital pathology systems. *J Pathol Inform*. 2013;4:21.
 75. Tellez D, van der Laak J, Ciompi F. Gigapixel whole-slide image classification using unsupervised image compression and contrastive training [conference paper]. In: 1st Conference on Medical Imaging with Deep Learning (MIDL 2018); 2018, 4–6th July; Amsterdam, The Netherlands, published online at <https://openreview.net/forum?id=Hk2YYqsf>.
 76. Williams BJ, Bottoms D, Treanor D. Future-proofing pathology: the case for clinical adoption of digital pathology. *J Clin Pathol*. 2017;70(12):1010–8. <https://doi.org/10.1136/jclinpath-2017-204644>.
 77. Yagi Y, Gilbertson JR. Digital imaging in pathology: the case for standardization. *J Telemed Telecare*. 2005;11(3):109–16. <https://doi.org/10.1258/1357633053688705>. PMID: 15901437.
 78. Herrmann MD, Clunie DA, Fedorov A, Doyle SW, Pieper S, Klepeis V, et al. Implementing the DICOM standard for digital pathology. *J Pathol Inform*. 2018;9:37. https://doi.org/10.4103/jpi.jpi_42_18.
 79. Treanor D, Gallas BD, Gavrielides MA, Hewitt SM. Evaluating whole slide imaging: a working group opportunity. *J Pathol Inform*. 2015;6:4.
 80. Gurcan MN, Boucheron LE, Can A, Madabhushi A, Rajpoot NM, Yener B. Histopathological image analysis: a review. *IEEE Rev Biomed Eng*. 2009;2:147–71.
 81. Hinton GE. A practical guide to training restricted Boltzmann machines. In: Neural networks: tricks of the trade. Berlin/New York: Springer; 2012. p. 599–619.
 82. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Adv Neural Inf Proces Syst*. 2012;25:1097–105.
 83. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–44. <https://doi.org/10.1038/nature14539>.

84. Corne SA. Artificial neural networks for pattern recognition. *Concepts Magn Reson.* 1996;8(5):303–24.
85. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal.* 2017;42:60–88.
86. Kazeminia S, Baur C, Kuijper A, van Ginneken B, Navab N, Albarqouni S, et al. GANs for medical image analysis. *Artif Intell Med.* 2020;109:101938.
87. Salvi M, Acharya UR, Molinari F, Meiburger KM. The impact of pre-and post-image processing techniques on deep learning frameworks: a comprehensive review for digital pathology image analysis. *Comput Biol Med.* 2020;128:104129.
88. Serre T. Deep learning: the good, the bad, and the ugly. *Ann Rev Vis Sci.* 2019;5:399–426.
89. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. European conference on computer vision. Springer; 2014. p. 818–833.
90. Gallas BD, Chan H-P, D’Orsi CJ, Dodd LE, Giger ML, Gur D, et al. Evaluating imaging and computer-aided detection and diagnosis devices at the FDA. *Acad Radiol.* 2012;19(4):463–77.
91. Mayer D, Butler D. Statistical validation. *Ecol Model.* 1993;68(1–2):21–32.
92. Meijering E. A bird’s-eye view of deep learning in bioimage analysis. *Comput Struct Biotechnol J.* 2020;18:2312.
93. Stadler CB, Lindvall M, Lundström C, Bodén A, Lindman K, Rose J, et al. Proactive construction of an annotated imaging database for artificial intelligence training. *J Digit Imaging.* 2020;34:1–11.
94. Tajbakhsh N, Jeyaseelan L, Li Q, Chiang JN, Wu Z, Ding X. Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation. *Med Image Anal.* 2020;63:101693.
95. Gray A, Wright A, Jackson P, Hale M, Treanor D. Quantification of histochemical stains using whole slide imaging: development of a method and demonstration of its usefulness in laboratory quality control. *J Clin Pathol.* 2015;68(3):192–9.
96. Clarke E, Revie C, Brettle D, Wilson R, Mello-Thoms C, Treanor D. Color calibration in digital pathology: the clinical impact of a novel test object [abstract]. In: 13th European Congress on Digital Pathology. 2016, May 25–28. Berlin, Germany. *Diagn Pathol.* 2016;1(8). Abstract P44.
97. Magee D, Treanor D, Crellin D, Shires M, Smith K, Mohee K, et al. Colour normalisation in digital histopathology images. Proc Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy (MICCAI Workshop). Citeseer; 2009. p. 100–11.
98. Wong NA, Hunt LP, Novelli MR, Shepherd NA, Warren BF. Observer agreement in the diagnosis of serrated polyps of the large bowel. *Histopathology.* 2009;55(1):63–6.
99. Parikh RB, Obermeyer Z, Navathe AS. Regulation of predictive analytics in medicine. *Science (New York, NY).* 2019;363(6429):810–2. <https://doi.org/10.1126/science.aaw0029>.
100. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *BMJ.* 2020;370:m3164.
101. Rivera SC, Liu X, Chan A-W, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *BMJ.* 2020;370:m3210.
102. Sounderahaj V, Ashrafiyan H, Aggarwal R, De Fauw J, Denniston AK, Greaves F, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: the STARD-AI Steering Group. *Nat Med.* 2020;26(6):807–8. <https://doi.org/10.1038/s41591-020-0941-1>.
103. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet.* 2019;393 (10181):1577–9. [https://doi.org/10.1016/S0140-6736\(19\)30037-6](https://doi.org/10.1016/S0140-6736(19)30037-6).
104. Moxley-Wyles B, Colling R, Verrill C. Artificial intelligence in pathology: an overview. *Diagn Histopathol.* 2020;26(11):513–20. <https://doi.org/10.1016/j.dmpdp.2020.08.004>.
105. Williams BJ, Lee J, Oien KA, Treanor D. Digital pathology access and usage in the UK: results from a national survey on behalf of the National Cancer Research Institute’s CM-Path initiative. *J Clin Pathol.* 2018;71(5):463–6. <https://doi.org/10.1136/jclinpath-2017-204808>.
106. Thorstenson S, Molin J, Lundström C. Implementation of large-scale routine diagnostics using whole slide imaging in Sweden: digital pathology experiences 2006–2013. *J Pathol Inform.* 2014;5 (1):14. <https://doi.org/10.4103/2153-3539.129452>.
107. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology – new tools for diagnosis and precision oncology. *Nat Rev Clin Oncol.* 2019;16(11):703–15. <https://doi.org/10.1038/s41571-019-0252-y>.
108. Colling R, Pitman H, Oien K, Rajpoot N, Macklin P, Snead D, et al. Artificial intelligence in digital pathology: a roadmap to routine use in clinical practice. *J Pathol.* 2019;249(2):143–50. <https://doi.org/10.1002/path.5310>.
109. Rakha EA, Toss M, Shiino S, Gamble P, Jaroensri R, Mermel CH, et al. Current and future applications of artificial intelligence in pathology: a clinical perspective. *J Clin Pathol.* 2020. <https://doi.org/10.1136/jclinpath-2020-206908>.
110. Allen TC. Regulating artificial intelligence for a successful pathology future. *Arch Pathol Lab Med.* 2019;143(10):1175–9. <https://doi.org/10.5858/arpa.2019-0229-ED>.
111. FDA. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD). Discussion paper and request for feedback; 2019. p. 1–20.

112. Ibex: Ibex Obtains CE-IVD Mark for AI-Powered Cancer Detection. <https://ibex-ai.com/press/ibex-obtains-ce-ivd-mark-for-ai-powered-cancer-detection/> (2020). Accessed 8 Dec 2020.
113. Roche: Roche improves speed and accuracy of non-small cell lung cancer diagnosis with launch of automated digital pathology algorithm. <https://www.roche.com/investors/updates/inv-update-2020-06-26.htm> (2020). Accessed 8 Dec 2020.
114. 4DPath: FDA Grants Breakthrough Designation to 4D Path for Novel Cancer Diagnostic Solution. <https://4dpath.com/fda-grants-breakthrough-designation-to-4d-path-for-novel-cancer-diagnostic-solution/> (2020). Accessed 8 Dec 2020.
115. Paige.AI: FDA Grants Breakthrough Designation to Paige.AI. <https://paige.ai/resources/fda-grants-breakthrough-designation-to-paige.ai> (2019). Accessed 8 Dec 2020.
116. Madabhushi A, Lee G. Image analysis and machine learning in digital pathology: challenges and opportunities. *Medical image analysis*. 2016;33:170–5.
117. Browning L, Colling R, Rakha E, Rajpoot N, Rittscher J, James JA, et al. Digital pathology and artificial intelligence will be key to supporting clinical and academic cellular pathology through COVID-19 and future crises: the PathLAKE consortium perspective. *J Clin Pathol*. 2020;jclinpath-2020-206854. <https://doi.org/10.1136/jclinpath-2020-206854>.



Artificial Intelligence in Kidney Pathology

38

Sato Noriaki, Uchino Eiichiro, and Okuno Yasushi

Contents

Introduction	540
Detection	541
Detection Using Conventional Features	541
Detection Using Deep Learning	542
Segmentation	542
Segmentation of Glomeruli	542
Segmentation of Multiple Structures	543
Classification	543
Classification of Major Pathological Findings	544
Classification and Identification of Specific Components	545
Classification Based on Pathological Category	545
Classification of Images with Immunohistochemistry	545
Classification Based on the Clinical Category and Genotype	546
Summary and Future Implications	546
Equations	547
References	547

Abstract

Recent breakthroughs in the classification of thousands of images into defined categories by deep learning has led to the application of these algorithms and models to the field of medicine, particularly in histology and pathology. The evaluation of histological images obtained by a kidney biopsy is a key step in the diagnosis and assessment of disease status in the practice of clinical nephrology. Many studies have investigated how deep learning can be applied to nephropathology to evaluate complex structures present inside the kidneys. In this chapter,

S. Noriaki · U. Eiichiro

Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto University, Kyoto, Japan

Department of Nephrology, Graduate School of Medicine, Kyoto University, Kyoto, Japan

O. Yasushi (✉)

Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto University, Kyoto, Japan
e-mail: okuno.yasushi.4c@kyoto-u.ac.jp

we provide a brief discussion on nephropathology and introduce some of the most recent studies on artificial intelligence involving this field. Additionally, we also discuss about the implications of and challenges associated with the application of AI to nephropathology. The studies are divided into three subcategories based on the key objective of the presented study. These include (i) identifying the position of specific structures, especially the glomerulus (detection), (ii) segmentation and identification of various structures present in histological images (segmentation), and (iii) classification of specific images into clinically and pathologically defined categories (classification).

Keywords

Kidney pathology · Machine learning · Artificial intelligence · Glomerulus · Deep learning · Renal pathology · Nephropathology

Introduction

The evaluation of the histological images obtained through a kidney needle biopsy specimen is a key step in the diagnosis and assessment of disease status in clinical nephrology, combined with demographic and clinical information such as laboratory data. The histological evaluation includes various structural components including glomeruli, urinary tubules, renal arteries, and cellular components as well as their morphology. The evaluation must be performed across multiple stains that are used routinely, each of which has its advantages and disadvantages, including the commonly used hematoxylin and eosin (H&E), periodic acid-Schiff (PAS), periodic acid-methenamine silver (PAM), and Masson's trichrome (MT). Additionally, immunofluorescence (IF) staining performed to assess the deposition of immunoglobulins and other proteins (including complement and light chains), along with findings of electron microscopy such as identification of dense deposits, are combined. Moreover, in the experimental investigation, histological images of a resected nephrectomy

specimen obtained from rats or mice need to be evaluated. These include the investigation of nephrotoxicity of medications or disease condition and progression in gene knockout mice.

Computational evaluation of histological images offers automation, quantifiability, and reproducibility and has historically been well described. A recent breakthrough in the classification of large image datasets into defined categories using machine learning (ML), especially deep learning (DL) owing to increasing computational power, advances in scanning technology, and the storage of large datasets of virtual slides in medical institutions, has led to the application of these algorithms and models to medicine, including histology and pathology. DL algorithms, especially neural network architectures, have the potential to extract deep features from images by learning through multiple layers and subsequently can be combined directly with other data such as clinical variables and high-throughput sequencing data to perform an integrated analysis, which could pave the way for personalized medicine.

The importance of artificial intelligence (AI), including ML and DL in the evaluation of nephropathology, is gradually being recognized by the nephrology community. Several international consortiums have gathered and organized renal biopsy images and the associated clinical metadata into digital pathology repositories, and the application of DL to the data stored in such repositories has already been published [1]. Additionally, obstacles such as inter-pathologist disagreement regarding assessment, which leads to incorrect labeling in supervised learning, have been recognized [2], and a descriptor of kidney histological images has been proposed for standardization [3]. Moreover, the Banff 2019 Meeting discussed the potential use of AI in organ transplantation, such as for an automated evaluation of Banff classification and development of personalized medicine [4]. Here, supervised learning refers to ML that has a pair of input and labeled output, contrary to unsupervised learning, such as a clustering analysis, where labeled output is not available, and the algorithm finds patterns in data without previous knowledge.

Most notably, the number of studies investigating the applications of DL for performing complicated tasks involved in the assessment of complex structures and heterogenic pathological features of kidney histopathological images has been steadily increasing. These investigations are being carried out by those who are specialized in ML, nephrology, and pathology. Thus, in this chapter, the current studies involving AI in nephropathology, with specific focus on application, techniques, and algorithms, are introduced and summarized as a tertiary source for readers from other fields. For clarity, the studies were divided into three subcategories according to the study's major objective of application: detection, segmentation, and classification. Additionally, the subcategories were further divided based on the investigated tasks.

Detection

The computational evaluation of the histological findings of glomeruli in whole slide images (WSI) has been a central focus in the field. Urine is generated in the renal corpuscle, which is a small sphere approximately 200 μm in diameter, composed of the glomerulus and Bowman's capsule and is present in the renal cortex. The glomeruli reflect a variety of pathological conditions. To assess the glomeruli computationally, one must first annotate the position of the glomeruli in the WSI; however, manual detection and position annotation of thousands of WSI are labor-intensive for both needle biopsies and kidney sections. Thus, computational methods to detect the glomeruli in the WSI have gained interest. The difficulty of the task lies in the fact that the glomeruli present in all the different stains should be evaluated; however, the position of the glomeruli differs across stains, and sometimes a glomerulus present in one staining is not available in others. The studies usually evaluate the detection performance using precision (Eq. 1), recall (Eq. 2), and F1 score, which is a weighted average of precision and recall. When calculating precision and recall, true positives are usually defined by defined thresholds of intersection over union, calculated

by the area of overlap and area of union between manually annotated structures and predicted regions.

Detection Using Conventional Features

Samsi et al. proposed a method to identify glomeruli present in H&E-stained mouse kidney biopsy images by simple color segmentation, identification of regions corresponding to Bowman's space, and subsequent grouping of objects belonging to the same glomeruli [5]. The approach involving hand-crafted or traditional image features and ML algorithms, such as support vector machines (SVMs), has been historically performed in many studies. SVM is a popular supervised learning algorithm for classification and regression tasks; it finds maximum-margin hyperplanes using support vectors. Kakimoto et al. detected glomeruli in desmin-stained histological images from rats using a histogram of oriented gradients (HOG), a descriptor representing the shape features calculated using the histograms of gradient orientation and SVM [6]. From the same group, Kato et al. developed and used a customized HOG (named segmental HOG) combined with SVM to detect glomeruli in the slide images stained with desmin from rats and obtained an average F1 score of 0.866 [7]. Simon et al. used local binary patterns (LBP) combined with SVM and DL to successfully extract glomeruli from images of multiple stainings for mice, rats, and humans. LBP corresponds to the image features obtained by calculating the relative value of each pixel compared to that of the surrounding neighboring pixels. They applied the method on specimens of diabetic nephropathy (DN), a kidney-related serious comorbidity of diabetes mellitus with pathologies such as diffuse and nodular glomerulosclerosis [8]. Marée et al. proposed an approach for detecting glomeruli that first detected luminal regions by image thresholding and subsequently fit an ellipse using Fitzgibbon's fitting method in kidney needle biopsy samples. They applied the approach to MT-stained slide images [9].

Detection Using Deep Learning

DL algorithms can extract unsupervised features, rather than a calculation based on defined mathematical formulas, which makes it good for application to unstructured image data, particularly complex images such as histological images. Recent studies used DL algorithms and features derived from DL, especially convolutional neural networks (CNNs), to determine the position of the glomerulus. CNN is used most commonly with tasks involving images. CNN is composed of characteristic layers, including convolutional layers, which shift filters through images and extract feature maps to learn the patterns in the images, and pooling layers that reduce the dimensionality of the output of the previous layer. In nephropathology, Temerinac-Ott et al. used CNN to detect glomeruli from needle biopsy and nephrectomy slide images by performing a mutual comparison of the classification results of multiple stains, thus improving detection performance. The F1 score for a PAS-stained slide of a needle biopsy was 0.828 [10]. Bukowy et al. located glomeruli positions in MT-stained whole-kidney section images of rats using region-based CNN (R-CNN) and obtained an average precision and recall of 96.9% and 96.8%. Additionally, their proposed method could assess glomerular depth profiles [11]. R-CNN is an approach that is made specifically for the task of object detection, which combines selective search and CNN [12].

The proposed approaches worked well for detecting glomeruli in WSI, with single staining or different types of staining, judging from the performance. This may aid in the rapid and automatic identification of the glomerulus position, which is essential for generating the training dataset used in the supervised learning task. Glomeruli detection approach is gradually being replaced by the segmentation approaches discussed below, which are capable of concurrent identification of other structures in kidney histological images.

Segmentation

Image segmentation refers to the partitioning of images into multiple meaningful objects and has been one of the central focuses of DL studies involving images, which has led to the invention of networks for specific purposes like biomedical imaging and the broad application in medical images. The kidney pathology involves not only the glomeruli but also structures like the proximal and distal urinary tubules and renal arteries, which are evaluated for conditions like urinary tubulitis or arteriosclerosis. Accurate segmentation and identification of these structures in kidney biopsy images are also important along with concurrent detection of glomeruli. For this purpose, DL has been extensively used and has been successful in segmenting normal structures and pathologically changed structures in the kidney.

Segmentation of Glomeruli

Although the primary purpose is segmentation, some studies mainly investigate glomeruli identification in the slide images using the segmentation approach. Gadermayr et al. investigated multiple problems regarding the detection and segmentation of kidney biopsy images. They proposed an approach to classify glomeruli and non-glomeruli with a small amount of training data [13], segment PAS-stained renal WSI without training data [14], and proposed a method to segment sparse small objects in WSI [15]. Most recently, they proposed an approach that combined stain-independent supervised segmentation and unsupervised segmentation to segment kidney histological images. They performed image-to-image transition using generative adversarial networks to segment staining with unlabeled data, using staining for which labeling is available. They successfully applied the approach to PAS, acid fuchsin orange G, collagen III, and CD31 stained slide images, which would aid in keeping the manual labor required for annotation to a minimum [16]. Additionally, although the study may fall into the classification subcategory, Bueno et al.

used SegNet and U-Net to perform semantic segmentation and classification of the patches of WSI into non-glomerulus, glomerulus, and sclerotic glomerulus [17]. U-Net is a fully convolutional network (FCN), a network that is composed of only convolutional layers, developed specifically for biomedical image segmentation [18]. SegNet is also one of the FCNs with encoder-decoder architecture [19]. The study group made the corresponding glomerulus dataset available from the public repository, which would enhance the validation of the future development of AI-related nephropathology [20].

Segmentation of Multiple Structures

The mainstream of the segmentation task is to segment multiple structures, and some studies have already assessed the performance of DL in this task. Jayapandian et al. segmented biopsy images of patients with minimal change disease, which is a pathological condition that exhibits little or no abnormality in the glomerulus, using DL models. They evaluated the performance of the model by ground-truth annotation of the glomerular tuft, glomerular unit, proximal tubular segment, distal tubular segment, peritubular capillaries, and arteries. The annotation was performed by five nephropathologists, with four major stains. They achieved a high F1 score for the structures in multiple stains and found that PAS staining performed the best [1]. Most recently, Bouteldja et al. segmented PAS-stained images of the kidneys of various species under multiple conditions, including healthy and pathologic, into regions with renal tubules, glomerular tuft, glomerulus, artery, artery lumen, vein, and other remaining tissues using a modified U-Net. They revealed that the model could successfully segment and quantify the structure inside the images and confirmed that the approach could be applied to humans; it has the potential to serve as the general segmentation model of kidney biopsy images. Interestingly, the model performed well in the murine models, which have distinct pathologic characteristics of various causes, indicating

potential for applicability in pathological and normal conditions [21]. Additionally, a recent study by Ginley et al. proposed the use of CNN to segment normal glomeruli and glomeruli afflicted by pathological conditions of glomerulosclerosis, interstitial fibrosis, and tubular atrophy; they used circling annotation provided by pathologists for PAS-stained images. They also validated the model's performance in other institutions [22].

These studies provided premises that the digital quantitative evaluation for major structures evaluated routinely in kidney biopsies is possible, especially in pathological conditions where diverse structural changes occur, in addition to normal conditions. However, the problem of the identification of other structures not evaluated or annotated by the models, or segmentation in other complex pathological conditions, needs to be investigated in future studies.

Classification

The glomerulus exhibits complex pathologies including mesangial hypercellularity, sclerosis caused by obstruction of the capillary space due to increased extracellular matrix, and crescent formation caused by extracapillary proliferation. In nephropathology, evaluation and classification of the findings present in an image of the glomerulus are crucial for evaluating slide images of patients with conditions such as suspected glomerulonephritis. This is because conditions such as crescent formation reflect high disease activity and poor prognosis. The classification and quantification of pathologies in structures other than the glomeruli have also been reported in the literature. Moreover, attempts to classify images according to clinical variables assessed in the clinical practice of nephrology have been described. Often combined with detection and segmentation discussed above, classification has been investigated extensively as these evaluations have potential clinical significance, such as for the automation of the staging of a specific disease. The classification performance of a defined label is often assessed by Cohen's kappa, which evaluates the inter-rater

agreement of categorical classification using the proportion of agreement and the expected proportion of agreement between raters.

Classification of Major Pathological Findings

The proposal for models or frameworks performing classification of having specific findings, often seen in clinical practice, has been well described in various stains. Barros et al. extracted features from glomeruli images using image processing and pattern recognition methods such as Otsu's thresholding, an automatic image thresholding method. They used the k-nearest neighbor algorithm, one of the ML algorithms that assign a label of the most general category near k objects in the feature space to classify the images as having proliferative glomerular lesions or not in H&E and PAS-stained slides. They obtained a precision and recall of 88% in the validation dataset [23]. Sheehan et al. proposed a framework for identifying glomeruli and scoring the mesangial matrix expansion in the PAS-stained slide images, using ML and some image processing methods. The scores obtained by the approach had a significant correlation with the manual scoring performed by renal pathologists [24]. Kannan et al. fine-tuned Inception V3 (the high-performance CNN architecture introduced in 2015 by Google regarding the classification of the images in ImageNet [25, 26]), to classify images obtained by the sliding window approach into the non-glomerulus, normal or partially sclerotic glomerulus, and globally sclerotic glomerulus categories within images stained with MT. They obtained a high accuracy of 92.7% in discriminating non-glomerulus images and images from other categories [27]. Fine-tuning here refers to making some of the layers including the last output layer of the pre-trained model and training the new model using its own datasets, thus transferring the knowledge learned in the other model. Marsh et al. proposed a method for the identification of globally sclerotic glomeruli from frozen section biopsies of kidney

transplantation donors using the fine-tuning of VGG16, a popular CNN architecture for ImageNet classification. The identification of the proportion of globally sclerotic glomeruli is important for assessing transplant outcomes, and the proposed approach yielded high performance [28]. For the evaluation taking multiple major stains into account, Wu et al. proposed a generator-to-classifier framework that classifies glomeruli into segmental, global sclerotic, and normal glomeruli. This approach considers features from the four major stains used routinely for evaluating kidney biopsy images generated by generative adversarial networks. They successfully classified sclerotic versus normal glomeruli and segmental and global sclerosis [29]. The studies mentioned above mainly investigated one or few of the findings evaluated in renal pathology. They succeeded in classifying important pathologies, especially global and segmental sclerosis; however, many other findings remained. For the simultaneous evaluation of models for multiple pathological findings, Uchino et al. developed classifiers of seven major pathological findings of a glomerulus, namely, global and segmental sclerosis, endocapillary proliferation, basement membrane structural change, mesangial matrix accumulation, mesangial cell proliferation, and crescent. By fine-tuning of Inception V3, they found that the classification performance of the models was close to that of nephrologists, although the performance was low to moderate, except for global sclerosis, in PAS and PAM staining [30]. Yamaguchi et al. evaluated the concordance of the annotation of 5 pathologists in 12 features and trained the CNN for features with high Cohen's kappa. The models discriminating the findings of capillary collapse and fibrous crescents yielded high performances, and the visualization results of the true positives showed that the CNN correctly pointed out the regions with respective features [31]. These studies proved that DL can be applied to discriminate between major pathological findings of glomeruli; however, more subtle findings have been poorly investigated or have low performance, possibly owing to a limited amount of available data.

Classification and Identification of Specific Components

The classification and identification of more specific and detailed components and pathologies have been investigated in several studies. These components have more inter-nephropathologist disagreements; therefore, the application of DL is considered necessary. Chagas et al. used CNN combined with SVM to classify glomeruli as having hypercellularity or not in H&E and PAS staining. They also further divided the hypercellularity into lesions of mesangial, endocapillary, and both regions and obtained an average accuracy of 82% [32]. Zeng et al. used CNN and image processing methods to locate glomeruli in the whole slide, identify glomerular lesions, and classify the glomeruli into segmental, global sclerotic, and crescentic in PAS-stained images of patients with IgA nephropathy (IgAN). Moreover, they divided the intrinsic cells into mesangial, endothelial, and podocyte types and compared the identification performance with that of junior pathologists, and the ground truth was given by senior pathologists. Additionally, network-calculated mesangial hypercellularity scores were compared with those provided by senior pathologists. The results indicated high Cohen's kappa values for the classification of segmental and global sclerotic and crescentic glomeruli and for mesangial score assessment. Additionally, the network identification of intrinsic cells inside glomeruli outperformed those of junior pathologists [33]. Chen et al. proposed a method to classify normal glomerulus and glomerulus having spike formation in membranous nephropathy (MN) in PAM-stained images using U-Net and multiple instance learning. Spike formation is one of the important changes induced by MN in the basement membrane [34]. Multiple instance learning is a weakly supervised learning method, applicable when a small amount of label is available [35]. These studies provided proof of concept that detailed features within glomeruli could be captured and quantified by DL.

Classification Based on Pathological Category

In some diseases, the pathological class or stage has been proposed and widely integrated into clinical guidelines used in daily practice. DL can carry out classification based on the annotated labels of these validated classes. For the classification of such classes and stages, Ginley et al. proposed a method to assess glomerular pathology in DN by combining CNN and recurrent neural networks (RNNs) in PAS staining. RNN is a neural network architecture that takes into account continuous information when learning, especially applicable in time series data or natural language processing. The classification of the DN stages determined by the approach had a moderate agreement with senior pathologists assessed by Cohen's kappa [36]. The same group proposed a method to automatically quantify glomerular lesions of lupus nephritis (LN) by a set of hand-crafted features and RNN. They classified biopsy images into class II to IV and additional V of LN and compared the results to pathologists. They achieved moderate Cohen's kappa [37]. Although these studies focused on DN and LN, other forms of glomerulonephritis that require staging, like the MEST-C score of IgAN, could be evaluated by the application of these models or other relevant DLs.

Classification of Images with Immunohistochemistry

In addition to stained light microscopic images, Ligabue et al. used CNN to classify images of IF staining. Assessment of IF staining is crucial in the diagnosis of conditions such as IgAN, and the classification of complex combinations of their appearance, distribution, location, and intensity requires expertise. They revealed that the performance of the models regarding classification was comparable with that of experienced pathologists, and the processing time of machines was faster than that of humans [38]. Choi et al. proposed a method combining Inception V3 and Faster R-CNN, a faster implementation of R-CNN, to

detect peritubular capillary (PTC) in the slide images and C4d immunohistological staining of PTC. This is an important histological feature of antibody-mediated rejection in kidney transplantation. The C4d score calculated by the proposed approach was in concordance with that of pathologists and had a significant relationship with clinical outcomes such as graft survival [39]. These applications in immunohistochemistry could be combined with other routine stains to achieve better classification of findings or prediction of renal outcome.

Classification Based on the Clinical Category and Genotype

One of the scarcely investigated tasks is the relationship between clinical variables and features derived from nephropathology images. One study investigated CNN performance on classification according to the clinical category. Kolachalama et al. assessed the relationship between histological images stained with MT and categorized clinical variables such as serum creatinine, proteinuria, and prognostic information corresponding to renal survival, by fine-tuning of Inception V3. The study revealed that for all the models of creatinine, proteinuria, and 1-, 3-, and 5-year renal survival, the CNN estimated better than the model based on the nephropathologist-estimated fibrosis score [40]. Additionally, combined with genetic information, Sheehan et al. used features derived from AlexNet, a popular CNN architecture, and SVM to detect glomeruli from WSI of PAS-stained mouse kidney histological images and classified the glomeruli based on the genotype. They successfully classified glomeruli into the genotype of Far2, which is related to mesangial matrix expansion and has been proved to increase in various clinical glomerulonephritis using features derived from glomerulus images. This study proved the potential of DL application to slide images to discriminate the genotype [41].

Summary and Future Implications

In this chapter, we summarized the current applications of DL in nephropathology, defining the subcategories of the investigated tasks. We introduced relevant studies and current perspectives for each subcategory. As summarized above, the segmentation of major components in kidney pathology has been performed in both pathological and normal slide images and successful. Additionally, many frameworks for classification across pathological findings and clinical variables have been described.

There are many other pathological conditions that are yet to be assessed, such as virus-associated kidney injury which is important in kidney transplantation; however, enough training data might not be available due to the rarity of these conditions. The performance of these domain-specific models is promising; however, the integration and implementation of the proposed frameworks in real-world clinical settings and the evaluation of the clinical significance of machine-derived classification and features, along with the interpretation of the rationale behind the prediction of DL models for human understanding, have not been satisfactorily described. This could be a future direction for the field. Additionally, the combination of features derived from established models of kidney pathology images, combined with transcriptomes, radiomics, or data from electronic health records for the prediction of outcomes or disease mechanisms, namely, multimodal learning, is still poorly investigated in the nephrology field compared to the fields like oncopathology. Most of the investigated tasks are performed by supervised learning, and extracting the features from kidney images in an unsupervised manner could extract novel features related to clinical outcomes.

DL application in nephropathology would aid in assisting nephrologists or pathologists clinically. The methods and algorithms of ML are still improving, and an in-depth understanding of the advantages and disadvantages of these newly invented methods is needed for

the implementation of these technologies in real clinical practice.

Equations

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}} \quad (1)$$

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}} \quad (2)$$

References

1. Jayapandian CP, Chen Y, Janowczyk AR, Palmer MB, Cassol CA, Sekulic M, et al. Development and evaluation of deep learning-based segmentation of histologic structures in the kidney cortex with multiple histologic stains. *Kidney Int.* 2021;99(1):86–101. <https://doi.org/10.1016/j.kint.2020.07.044>.
2. Bellur SS, Roberts ISD, Troyanov S, Royal V, Coppo R, Cook HT, et al. Reproducibility of the Oxford classification of immunoglobulin A nephropathy, impact of biopsy scoring on treatment allocation and clinical relevance of disagreements: evidence from the VALIDation of IGA study cohort. *Nephrol Dial Transplant.* 2019;34(10):1681–90. <https://doi.org/10.1093/ndt/gfyz337>.
3. Barisoni L, Troost JP, Nast C, Bagnasco S, Avila-Casado C, Hodgin J, et al. Reproducibility of the NEPTUNE descriptor-based scoring system on whole-slide images and histologic and ultrastructural digital images. *Mod Pathol.* 2016;29(7):671–84. <https://doi.org/10.1038/modpathol.2016.58>.
4. Loupy A, Haas M, Roufosse C, Naesens M, Adam B, Afrouzian M, et al. The Banff 2019 Kidney Meeting Report (I): Updates on and clarification of criteria for T cell- and antibody-mediated rejection. *Am J Transplant.* 2020;20(9):2318–31. <https://doi.org/10.1111/ajt.15898>.
5. Sami S, Jarjour WN, Krishnamurthy A. Glomeruli segmentation in H&E stained tissue using perceptual organization. In: 2012 IEEE Signal Processing in Medicine and Biology Symposium (SPMB). 2012. p. 1–5. <https://doi.org/10.1109/SPMB.2012.6469464>.
6. Kakimoto T, Okada K, Hirohashi Y, Relator R, Kawai M, Iguchi T, et al. Automated image analysis of a glomerular injury marker desmin in spontaneously diabetic Torii rats treated with losartan. *J Endocrinol.* 2014;222(1):43–51. <https://doi.org/10.1530/JOE-14-0164>.
7. Kato T, Relator R, Ngou H, Hirohashi Y, Takaki O, Kakimoto T, et al. Segmental HOG: new descriptor for glomerulus detection in kidney microscopy image. *BMC Bioinformatics.* 2015;16:316. <https://doi.org/10.1186/s12859-015-0739-1>.
8. Simon O, Yacoub R, Jain S, Tomaszewski JE, Sarder P. Multi-radial LBP Features as a Tool for Rapid Glomerular Detection and Assessment in Whole Slide Histopathology Images. *Sci Rep.* 2018;8(1):2032. <https://doi.org/10.1038/s41598-018-20453-7>.
9. Marée R, Dallongeville S, Olivo-Marin J, Meas-Yedid V. An approach for detection of glomeruli in multisite digital pathology. In: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI). 2016. p. 1033–6. <https://doi.org/10.1109/ISBI.2016.7493442>.
10. Temerinac-Ott M, Forestier G, Schmitz J, HermSEN M, Bräsen JH, Feuerhake F, et al. Detection of glomeruli in renal pathology by mutual comparison of multiple staining modalities. In: Proceedings of the 10th International Symposium on Image and Signal Processing and Analysis. 2017. p. 19–24. <https://doi.org/10.1109/ISPA.2017.8073562>.
11. Bukowy JD, Dayton A, Cloutier D, Manis AD, Staruschenko A, Lombard JH, et al. Region-based convolutional neural nets for localization of glomeruli in trichrome-stained whole kidney sections. *J Am Soc Nephrol.* 2018;29(8):2081–8. <https://doi.org/10.1681/ASN.2017111210>.
12. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. 2014. p. 580–7. <https://doi.org/10.1109/CVPR.2014.81>.
13. Gadermayr M, Klinkhammer BM, Boor P, Merhof D. Do we need large annotated training data for detection applications in biomedical imaging? A case study in renal glomeruli detection. In: machine learning in medical imaging. Springer International Publishing; 2016. p. 18–26. https://doi.org/10.1007/978-3-319-47157-0_3.
14. Gadermayr M, Eschweiler D, Jeevanesan A, Klinkhammer BM, Boor P, Merhof D. Segmenting renal whole slide images virtually without training data. *Comput Biol Med.* 2017;90:88–97. <https://doi.org/10.1016/j.combiomed.2017.09.014>.
15. Gadermayr M, Dombrowski A-K, Klinkhammer BM, Boor P, Merhof D. CNN cascades for segmenting sparse objects in gigapixel whole slide images. *Comput Med Imaging Graph.* 2019;71:40–8. <https://doi.org/10.1016/j.compmedimag.2018.11.002>.
16. Gadermayr M, Gupta L, Appel V, Boor P, Klinkhammer BM, Merhof D. Generative adversarial networks for facilitating stain-independent supervised and unsupervised segmentation: a study on kidney histology. *IEEE Trans Med Imaging.* 2019;38(10):2293–302. <https://doi.org/10.1109/TMI.2019.2899364>.

17. Bueno G, Fernandez-Carrobles MM, Gonzalez-Lopez L, Deniz O. Glomerulosclerosis identification in whole slide images using semantic segmentation. *Comput Methods Programs Biomed.* 2020;184:105273. <https://doi.org/10.1016/j.cmpb.2019.105273>.
18. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: *Medical image computing and computer-assisted intervention – MICCAI 2015*. Springer International Publishing; 2015. p. 234–41. https://doi.org/10.1007/978-3-319-24574-4_28.
19. Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal Mach Intell.* 2017;39(12):2481–95. <https://doi.org/10.1109/TPAMI.2016.2644615>.
20. Bueno G, Gonzalez-Lopez L, Garcia-Rojo M, Laurinavicius A, Deniz O. Data for glomeruli characterization in histopathological images. *Data Brief.* 2020;29:105314. <https://doi.org/10.1016/j.dib.2020.105314>.
21. Bouteldja N, Klinkhammer BM, Bülow RD, Drost P, Otten SW, von Stillfried SF, et al. Deep learning-based segmentation and quantification in experimental kidney histopathology. *J Am Soc Nephrol.* 2021;32(1):52–68. <https://doi.org/10.1681/ASN.2020050597>.
22. Ginley B, Jen K-Y, Han SS, Rodrigues L, Jain S, Fogó AB, et al. Automated computational detection of interstitial fibrosis, tubular atrophy, and glomerulosclerosis. *J Am Soc Nephrol.* 2021;32(4):837–50. <https://doi.org/10.1681/ASN.2020050652>.
23. Barros GO, Navarro B, Duarte A, Dos-Santos WLC. PathoSpotter-K: a computational tool for the automatic identification of glomerular lesions in histological images of kidneys. *Sci Rep.* 2017;7:46769. <https://doi.org/10.1038/srep46769>.
24. Sheehan SM, Korstanje R. Automatic glomerular identification and quantification of histological phenotypes using image analysis and machine learning. *Am J Physiol Renal Physiol.* 2018;315(6):F1644–51. <https://doi.org/10.1152/ajprenal.00629.2017>.
25. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis.* 2015;115(3):211–52. <https://doi.org/10.1007/s11263-015-0816-y>.
26. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016. p. 2818–26. <https://doi.org/10.1109/CVPR.2016.308>.
27. Kannan S, Morgan LA, Liang B, Cheung MG, Lin CQ, Mun D, et al. Segmentation of glomeruli within trichrome images using deep learning. *Kidney Int Rep.* 2019;4(7):955–62. <https://doi.org/10.1016/j.kir.2019.04.008>.
28. Marsh JN, Matlock MK, Kudose S, Liu T-C, Stappenbeck TS, Gaut JP, et al. Deep learning global glomerulosclerosis in transplant kidney frozen sections. *IEEE Trans Med Imaging.* 2018;37(12):2718–28. <https://doi.org/10.1109/TMI.2018.2851150>.
29. Wu B, Zhang X, Zhao S, Xie L, Zeng C, Liu Z, et al. G2C: a generator-to-classifier framework integrating multi-stained visual cues for pathological glomerulus classification. *Proceedings of the AAAI Conference on Artificial Intelligence.* 2019;33(01):1214–21. <https://doi.org/10.1609/aaai.v33i01.33011214>.
30. Uchino E, Suzuki K, Sato N, Kojima R, Tamada Y, Hiragi S, et al. Classification of glomerular pathological findings using deep learning and nephrologist-AI collective intelligence approach. *Int J Med Inform.* 2020;141:104231. <https://doi.org/10.1016/j.ijmedinf.2020.104231>.
31. Yamaguchi R, Kawazoe Y, Shimamoto K, Shinohara E, Tsukamoto T, Shintani-Domoto Y, et al. Glomerular classification using convolutional neural networks based on defined annotation criteria and concordance evaluation among clinicians. *Kidney Int Rep.* 2021;6(3):716–26. <https://doi.org/10.1016/j.ekir.2020.11.037>.
32. Chagas P, Souza L, Araújo I, Aldeman N, Duarte A, Angelo M, et al. Classification of glomerular hypercellularity using convolutional features and support vector machine. *Artif Intell Med.* 2020;103:101808. <https://doi.org/10.1016/j.artmed.2020.101808>.
33. Zeng C, Nan Y, Xu F, Lei Q, Li F, Chen T, et al. Identification of glomerular lesions and intrinsic glomerular cell types in kidney diseases via deep learning. *J Pathol.* 2020;252(1):53–64. <https://doi.org/10.1002/path.5491>.
34. Chen Y, Li M, Hao F, Han W, Niu D, Wang C. Classification of glomerular spikes using Convolutional Neural Network. In: *Proceedings of the 2020 Conference on Artificial Intelligence and Healthcare*. New York, NY, USA: Association for Computing Machinery; 2020. p. 254–8. <https://doi.org/10.1145/3433996.3434043>.
35. Dietterich TG, Lathrop RH, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles. *Artif Intell.* 1997;89(1):31–71. [https://doi.org/10.1016/S0004-3702\(96\)00034-3](https://doi.org/10.1016/S0004-3702(96)00034-3).
36. Ginley B, Lutnick B, Jen K-Y, Fogó AB, Jain S, Rosenberg A, et al. Computational segmentation and classification of diabetic glomerulosclerosis. *J Am Soc Nephrol.* 2019 Oct;30(10):1953–67. <https://doi.org/10.1681/ASN.2018121259>.
37. Ginley B, Jen K-Y, Rosenberg A, Rossi GM, Jain S, Sarder P. Fully automated classification of glomerular lesions in lupus nephritis. In: *Medical Imaging 2020: Digital Pathology*. International Society for Optics and Photonics; 2020. p. 113200Y. <https://doi.org/10.1117/12.2548528>.
38. Ligabue G, Pollastri F, Fontana F, Leonelli M, Furci L, Giovanella S, et al. Evaluation of the classification accuracy of the kidney biopsy direct immunofluorescence through convolutional neural networks. *Clin J Am Soc Nephrol.* 2020;15(10):1445–54. <https://doi.org/10.2215/CJN.03210320>.

39. Choi G, Kim Y-G, Cho H, Kim N, Lee H, Moon KC, et al. Automated detection algorithm for C4d immunostaining showed comparable diagnostic performance to pathologists in renal allograft biopsy. *Mod Pathol.* 2020;33(8):1626–34. <https://doi.org/10.1038/s41379-020-0529-9>.
40. Kolachalama VB, Singh P, Lin CQ, Mun D, Belghasem ME, Henderson JM, et al. Association of pathological fibrosis with renal survival using deep neural networks. *Kidney Int Rep.* 2018;3(2):464–75. <https://doi.org/10.1016/j.kir.2017.11.002>.
41. Sheehan S, Mawe S, Cianciolo RE, Korstanje R, Mahoney JM. Detection and classification of novel renal histologic phenotypes using deep neural networks. *Am J Pathol.* 2019;189(9):1786–96. <https://doi.org/10.1016/j.ajpath.2019.05.019>.



Christian Greis

Contents

Introduction	552
AIM in Dermatology	552
Area of Application: Skin Cancer	552
Area of Application: Psoriasis	554
Area of Application: Eczema	555
Area of Application: Other Skin Disorders	556
Dermatologist Attitude Toward Artificial Intelligence	556
Limitation of Artificial Intelligence in Dermatology	556
Better Applicability of AI Thanks to Teledermatology	557
Ethnic Variations as a Challenge in the Development of Algorithms	557
References	558

Abstract

In medicine, dermatology is a promising pioneer for the use of artificial intelligence (AI). In dermatological practice, the recognition of visual patterns (morphology) has always been fundamental for making a diagnosis, so artificial intelligence has great potential here. The collection of clinical data, especially image data, is playing an increasingly important role in the diagnosis and therapy of skin disease. Existing analog data archives are being digitized and restructured with great effort, and

new data sets are often captured and labeled directly in digital form. During the last years, a growing number of studies have demonstrated AI's benefits in research settings, and first applications are already used clinically. Particularly in the detection of skin cancer and for the quantification of chronic inflammatory skin diseases, artificial intelligence is supporting doctors as well as patients to find the right diagnosis and treatment. The purpose of this book chapter is to discuss the potential applications of artificial intelligence in various areas of dermatology.

Keywords

Dermatology · Digitization · Machine learning · Artificial intelligence · Deep learning · Convolutional neural network · Image recognition · Pattern recognition ·

C. Greis (✉)
Department of Dermatology, University Hospital Zurich,
Zurich, Switzerland
e-mail: christian.greis@usz.ch;
christian.greis@derma2go.com

Algorithm-based decision-making · Skin cancer · Psoriasis · Eczema · Teledermatology · Telemedicine

Conflict of Interest Dr. C. Greis is founder of the teledermatology platform www.darma2go.com.

Introduction

Up to two billion people worldwide are affected by skin conditions, prevalent across locations and age groups. Dermatology is a multifaceted discipline, ranging from aesthetic complaints to chronic inflammatory diseases and malignoma. Skin cancer is the most common cancer globally, with melanoma being the deadliest form.

Dermatology is a pioneer for the use of artificial intelligence (AI) in the field of medicine. The recognition of visual patterns (morphology) is fundamental for making a diagnosis in the dermatological practice, so artificial intelligence has great potential here. The collection of image data is playing an increasingly important role, focusing on diagnosis and treatment of skin disease. Existing analog data archives are digitized and structured with great effort, and new data sets are often captured and labeled directly in digital form. During the last years, a growing number of studies have demonstrated its benefits in research settings, and first applications are already in clinical usage. Particularly in the detection of skin cancer and for the quantification of chronic inflammatory skin diseases, artificial intelligence is supporting doctors as well as patients to find the right diagnose and establish sufficient treatment. The purpose of this chapter is to better understand the application possibilities of artificial intelligence in various fields of dermatology.

AIM in Dermatology

In the later part of the 2010s, artificial intelligence has gradually provided increasing clinical value in different areas of dermatology including skin cancer, psoriasis, eczema, and other skin diseases. In January 2017, a publication by Stanford

University in Nature attracted a great deal of media attention. A neural network achieved sensitivity and specificity comparable to 21 human dermatologists in distinguishing between keratinocyte carcinomas and benign seborrheic keratoses as well as between malignant melanomas and benign nevi on 130,000 histologically verified clinical test images [1]. Three years later, another study in Nature presented similar results for differential diagnosis across the 26 most common skin conditions in primary care, which for the first time analyzed images in combination with medical history data [2]. The basis for these studies is the so-called convolutional neural networks, which analyze image data in a manner inspired by biological processes of receptive fields. From the color values per pixel, simple structures called “features” are first extracted and combined into increasingly complex features, which should ultimately enable the recognition of an object category. Mathematical filter operations form the basis of this feature calculation. These processes can be used to highlight lines, edges, or other basic structures. These processes take place in an increasingly automated manner through pattern recognition and improve independently with an increasing amount of supplied image data.

Area of Application: Skin Cancer

Especially in image-based skin cancer diagnosis, computer-based image analysis has become increasingly important. Skin cancer in general is one of the most widespread cancers worldwide, with nonmelanoma skin cancer the most common cancer with a lifetime prevalence in Central and Northern Europe of 10–15% and increasing incidence [3]. Early detection of skin cancer has a major impact on prognosis. In the detection of skin cancer and especially in the differentiation of moles from melanoma, dermoscopy plays a central role, complementary to macroscopic assessment and medical history. Correct dermoscopic diagnosis is often a major challenge, as a broad spectrum of rarer differential diagnoses must first be excluded.

The analysis of dermoscopic images for the evaluation of moles and melanomas has been described in several research papers. In recent studies, convolutional neural networks outperformed dermatologists in distinguishing dermoscopic images of melanoma and nevi. A review from 2019 counted 1694 performed studies in this field, the majority in the field of computer science [4]. The widely cited study “Man against machine” from Heidelberg, Germany, investigated the diagnostic precision of a trained convolutional neural network compared to dermatologists and experts from all over the world and found that it was superior to most but not all dermatologists [5]. Based on this study, an algorithm-based evaluation tool for melanocytic lesions was approved for clinical use, which analyzes a dermoscopic image of a lesion and generates a numerical value that classifies the lesion as malignant or benign. In another study, the convolutional neural network also performed better than the dermatologists in most cases, in terms of the dermatopathological results of the lesions [6]. Nevertheless, it should be noted that these results were described under controlled conditions and cannot be replicated in clinical practice to date. Also, in the analysis of dermoscopic images of nonpigmented skin lesions, there are first publications comparing the diagnostic accuracy of

neural networks with that of dermatologists – with comparable results for both groups [7, 8]. With similar intention, a yearly event is conducted by the *International Skin Imaging Collaboration* (ISIC) [9]. Using a data set of over 25,000 dermoscopic images, international teams compete against each other with their algorithms. In all these projects and studies, dermatologists and artificial intelligence were considered as opponents. However, studies have found that the combination of human and artificial intelligence can achieve superior results to the independent results of either system [10] (Fig. 1).

Computer-based analysis of dermatoscopic images has the advantage that the images are usually acquired in a standardized fashion. This is different than when clinical, macroscopic images are acquired – especially when they are taken directly by the patients. Several applications (e.g., *SkinVision BV*) nowadays provide risk measures of pigmented and nonpigmented skin lesions based on macroscopic images taken by the patient. Some of these applications exhibit a high sensitivity in detecting skin cancer [11]. However, there is still room for improvement [12].

Additionally, to the analysis of dermoscopic and clinical images of melanoma, recent studies performed computer-based decision-making on dermatopathological images. For melanoma, the

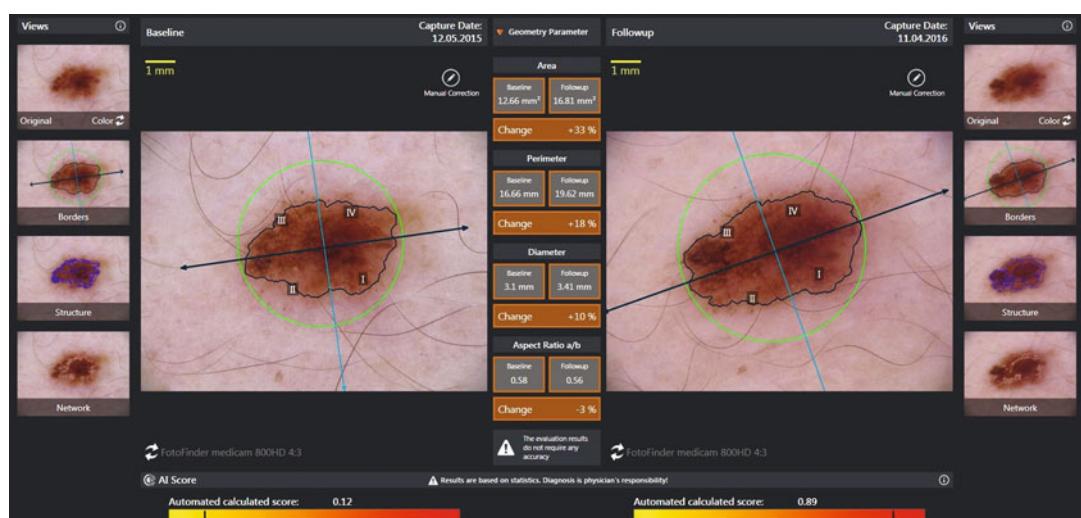


Fig. 1 Algorithm-based evaluation tool for melanocytic lesions. (© 2021 FotoFinder Systems GmbH)

literature reports 25–26% of discordance for classifying a benign nevus versus malignant melanoma between individual pathologists. Convolutional neural networks can thus be a valuable tool to assist human dermatopathological melanoma diagnoses [13].

Area of Application: Psoriasis

In addition to extensive studies on the detection of skin cancer using AI, the quantification as well as qualification of inflammatory dermatoses by computer-based decision-making has become a research topic with increasing popularity. With a prevalence of 2.5%, psoriasis is one of the most common inflammatory skin diseases among Caucasians.

The *Psoriasis Area and Severity Index* (PASI) and *body surface area* (BSA) criteria represent the gold standards for psoriasis severity assessments. A PASI and BSA of >10% or more indicate moderate to severe psoriasis, which is also the threshold for reimbursement of expensive systemic therapies in most countries.

Compared to examiner-dependent variability in PASI and BSA, AI-based algorithms have the potential for reproducible, standardized evaluations of these scores. Different studies show that

the assessment of erythema [14], as well as induration and scaling [15], is already technically possible through artificial intelligence. Computer-assisted programs for PASI can accurately calculate the proportion of psoriatic skin surface as well as the severity of erythema, induration, and desquamation by anatomic region. In pilot clinical validations, the technology showed a high reproducibility and high levels of agreement to results attained by PASI-trained physicians [16, 17] (Fig. 2).

Meanwhile, in addition to disease quantification, artificial intelligence can also help to provide a diagnosis of psoriasis. An algorithm that had to distinguish between 9 diagnoses based on 100 clinical pictures made fewer misdiagnoses of psoriasis and missed fewer diagnoses of psoriasis compared to 25 Chinese dermatologists [19].

After the diagnosis of psoriasis has been confirmed, artificial intelligence offers additional benefits in therapy planning. In a real-world setting, many patients with psoriasis need dose adjustments or changes to prescribed medication. One research project used the example of a biological therapy (IL-17A inhibitors) to show that machine learning can be used to determine which patients would benefit from modifications to the initial prescribed drug doses [20]. Another research team was able to predict the long-term



Fig. 2 Patient with psoriasis vulgaris. (a) Affected body surface area predicted by machine learning algorithm compared to manually marked area by experts. (b) Clinical findings with erythrosquamous plaques of the same patient

without markings [18]. (© Reprinted by permission from Springer, *Der Hautarzt*, Role of artificial intelligence in assessing the extent and progression of dermatoses, Maul et al. (2020))

responses of patients with psoriasis to biological therapies in general [21]. For the two biologics *tofacitinib* and *etanercept*, the long-term response to the drug treatment could be predicted by analyzing systemic inflammatory proteins measured before and several weeks after treatment initiation using a machine learning-based algorithm [22].

Overall, an improvement in diagnosis and treatment is expected in psoriasis patients, thanks to artificial intelligence.

Area of Application: Eczema

In contrast to the skin changes of psoriasis, which are mostly sharply defined and with a high contrast to unaffected skin, eczema phenotypes are often diffuse and vary depending on the disease stage and the cause of the eczema. For computer-assisted image diagnosis of eczema diseases, the challenge therefore is not only to discriminate correctly between healthy and affected skin but also to differentiate between different forms of eczema. In order to train algorithms to provide an AI-supported image analysis of these diverse assessment parameters, a large initial quantity of

image files is required. These might need to be further processed or enhanced.

Thus, current literature only contains sparse work on the use of artificial intelligence in the context of eczema. There are studies detecting eczema versus noneczema cases [23] or evaluating electronic patient data for phenotyping potential atopic patients, with predictive mechanistic models of progression and treatment response [24]. In studies with more clinical relevance, algorithms could identify various forms of eczema (e.g., seborrheic dermatitis, chronic eczema) [25] and were able to differentiate between eczematous and infectious conditions [26]. The latter distinction is particularly interesting, as it could be used to derive therapeutic suggestions.

Experimental studies have also investigated the detection of hand eczema from standardized images. However, clinical applicability in daily practice has not yet been achieved [27]. For example, jewelry or deep hand lines were misinterpreted as diseased skin. Also, the detection of less pronounced eczema is currently still problematic (Fig. 3).

Overall, algorithm-based analysis for an examiner-independent diagnosis of the spread

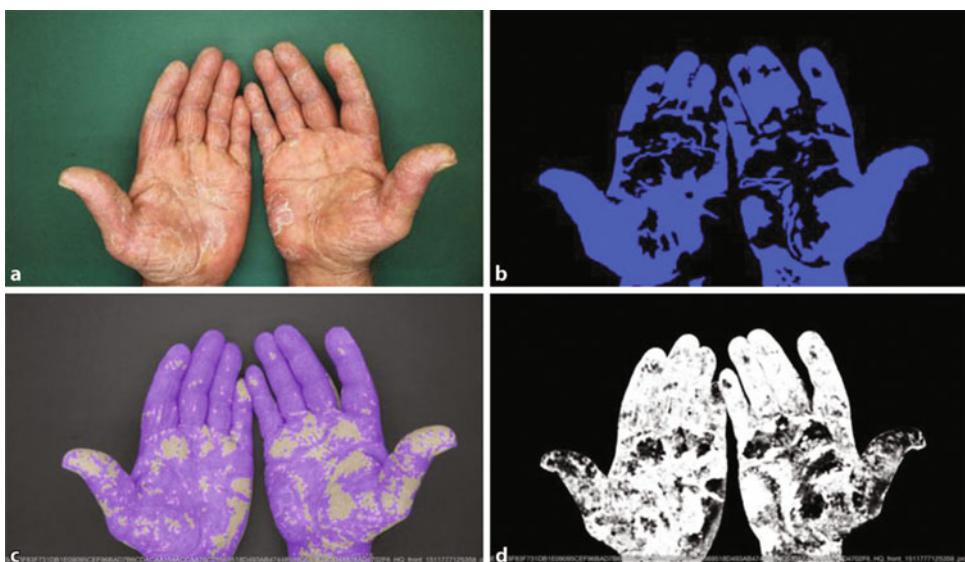


Fig. 3 Performance of the trained algorithm in a case with hand eczema. (a) Input image. (b) Extracted expert label. (c, d) Superposition of the computed images [18]. (©

Reprinted by permission from Springer, *Der Hautarzt*, Role of artificial intelligence in assessing the extent and progression of dermatoses, Maul et al. (2020))

and severity of eczema disease that is comparable to diagnoses existing for psoriasis is not yet available.

Area of Application: Other Skin Disorders

There is also preliminary experimental experience with the use of artificial intelligence in other dermatological diseases. In the diagnosis and quantification of vitiligo [18], acne vulgaris [28], onychomycosis [29], allergology [30], and dermatopathology [13], algorithm-based diagnostics performed similarly to comparison groups of dermatologists.

An early 2017 study demonstrated the possibility of diagnosing acne vulgaris using a neural network [31]. Recent publications on the application of computer-assisted systems to acne vulgaris are based on large data sets of cell phone photographs of affected patients. The data sets are used to develop AI-based algorithms to determine the severity of facial acne and identify different types of acne lesions or post-inflammatory hyperpigmentation [32]. Other applications monitor the disease progression and recommend a visit to the doctor if necessary. Especially in the case of acne vulgaris, such a procedure could prevent more severe consequences with permanent skin damage in the form of scars.

Recently, artificial intelligence has also been successfully applied to the diagnosis of onychomycosis. Based on almost 50,000 finger and toenail images, the computer-based image analysis was able to demonstrate a higher diagnostic accuracy than a group of dermatologists [29].

These new technologies have also found increasing application in the beauty industry. There are the so-called smart mirror analyzers on the Internet (e.g., *HAUT.AI*), which are AI-supported technologies with image recognition systems that analyze the skin and are supposed to recommend skin care products based on the appearance of the skin and the current weather forecast. It remains questionable to what extent such applications are misused for advertising purposes by cosmetic professionals and whether they

dissuade patients from consulting a doctor when necessary.

Dermatologist Attitude Toward Artificial Intelligence

Interestingly, although the number of studies is increasing, there are relatively few papers in which dermatologists are significantly involved in the conception, design, and interpretation of the studies. Most studies are led by computer scientists and engineers [33]. As a practicing dermatologist, one must become aware of the potential benefits, but also the possible disadvantages in daily clinical routine. Overall, dermatologists have an optimistic attitude toward artificial intelligence. Majority of dermatologists believe that artificial intelligence will improve dermatology and think that it should be a part of medical training. An increasing understanding of the application of artificial intelligence within dermatology was correlated with a positive attitude. Only a minority of dermatologists agree or strongly agree that the human dermatologist will be replaced by artificial intelligence in the foreseeable future [34]. Ultimately, it is wise to consider technological advances in a healthcare system while critically analyzing the drawbacks of such innovations.

Limitation of Artificial Intelligence in Dermatology

Time and again, technical systems surpass human performance under laboratory conditions and then fail in everyday clinical practice due to aspects as process integration, acceptance by users, inoperability in stressful situations, or the technical circumstances. High significance in research under experimental conditions does not correlate with the actual significance in practice.

Even though the use of AI in outpatient settings, such as for dermoscopic differentiation of malignant and benign skin tumors, seems to become increasingly realistic, there are several limitations to the usage of artificial intelligence

in daily dermatological practice. Available and relevant medical image data can be limited, due to rarely occurring conditions, privacy issues, and especially the lack of human experts available to annotate training data. Also, the quality of image data is often lacking due to overexposure or underexposure, reflection, or blurred pictures. Additional misdiagnoses often occur due to artifacts (e.g., tattoos, hairs, air bubbles, skin flakes, markings, coffee stains) [35]. The accuracy in differentiation that can be achieved by machine learning depends largely on the quantity as well as the quality of the available data.

Besides the technical challenges that limit the application of artificial intelligence in clinical practice, the lack of public acceptance also plays a decisive role. This is significantly lower in medicine than in other areas (e.g., smartphone facial recognition). In addition to data protection, a lack of patient understanding also seems to limit willingness to use AI-based services.

Better Applicability of AI Thanks to Teledermatology

National and international working groups support the establishment of teledermatological projects (e.g., *derma2go AG* [36], *e-derm-consult GmbH* [37]) and in recent years have increasingly been using AI-based technologies to support doctors on site. Previous efforts to establish AI in everyday dermatology have had limited success because of the lack of concrete use cases in clinical practice and the fact that algorithms had been developed using standardized images. Recent learning success in automated diagnosis of skin lesions using clinical smartphone images, some of which are low-resolution, suggests that these technologies are now ready to be tested in or made available to daily practice [38]. Thus, AI in telemedicine is becoming an achievable reality to expand dermatologic care options and further reduce costs and waiting times. For example, AI could serve as a clinical diagnostic support tool for dermatologists engaging in telemedicine or to direct ambiguous or malignant pathologies to an available physician. Solutions are needed which are also viable for a

large number of patients. These must focus on being able to treat the simple problems as automatically as possible, so that the expertise of the few human dermatologists can be reserved for the evaluation and treatment of complex problems. As an example, the *PASSION project* aims to use AI to identify five visually easily identifiable dermatoses in children and treat them according to a standardized procedure [39]. This will not lead to the elimination of in-person dermatology consultations, as not all dermatoses can be identified and treated using teledermatology. The best results were obtained when AI and physician were combined, especially when the physician was inexperienced. Because of a shortage of dermatologists, cases are often seen by general practitioners with less diagnostic experience (Fig. 4).

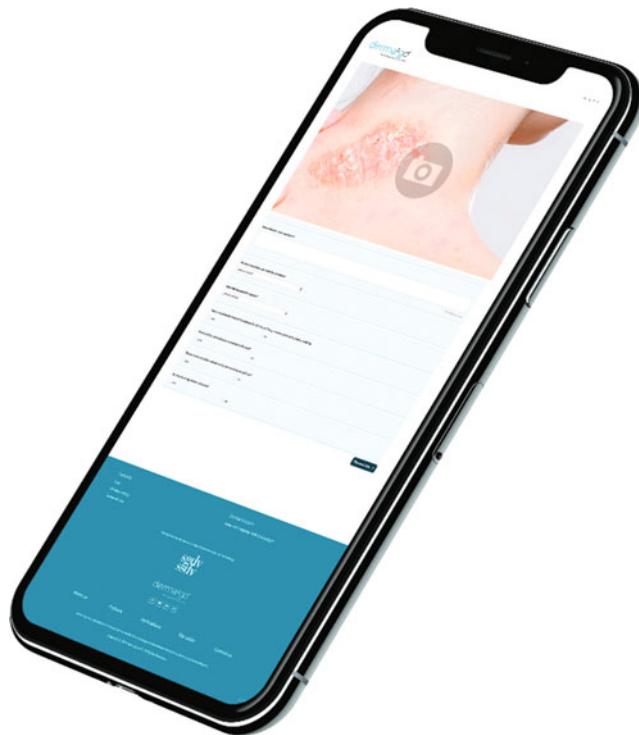
Ethnic Variations as a Challenge in the Development of Algorithms

Ethnic variations represent a challenge in the development of automated algorithms. Many algorithms developed so far were based on clinical images of Caucasians with skin types I–III according to Fitzpatrick. Conventional clinical experience shows that pathologies may present differently in their efflorescences (especially redness, scaling, lichenification) depending on the skin type. Also, dermatologic pathologies, such as the manifestations of atopic dermatitis, differ worldwide. In Africa, patients with atopic dermatitis have more follicular and lichenoid lesions; in Asia, lichenification and prurigo forms are more common; in India and Iran, flexural involvement is less frequent [40].

A pilot study in which an AI model was trained in an Asian population and subsequently validated in a Caucasian population found the procedure to be error-prone [41]. To further improve the accuracy of the systems in this regard and to be able to generalize to other ethnicities and skin types, it is important to increase the number of available clinical images of patients of different ages and ethnicities.

Several research groups worldwide have taken up this topic. *The Africa Teledermatology Project*,

Fig. 4 AI in telemedicine is becoming an achievable aim. (© derma2go AG)



supported by the *Commission for Development Studies of the Austrian Academy of Sciences* and the *American Academy of Dermatology*, has been building up a corresponding database for teaching and research purposes for years.

To further improve the accuracy of the systems and to be able to provide global solutions, it is important to increase the amount of available clinical data. This can only be achieved with the active participation of local healthcare providers as well as the dermatological community, always considering the interest of the individual patients.

References

1. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115–8. <https://doi.org/10.1038/nature21056>.
2. Liu Y, Jain A, Eng C, et al. A deep learning system for differential diagnosis of skin diseases. *Nat Med*. 2020;26(6):900–8. <https://doi.org/10.1038/s41591-020-0842-3>.
3. Chahal HS, Rieger KE, Sarin KY. Incidence ratio of basal cell carcinoma to squamous cell carcinoma equals with age. *J Am Acad Dermatol*. 2017;76(2):353–4. <https://doi.org/10.1016/j.jaad.2016.08.019>.
4. Dick V, Sinz C, Mittlbock M, et al. Accuracy of computer-aided diagnosis of melanoma: a meta-analysis. *JAMA Dermatol*. 2019; <https://doi.org/10.1001/jamadermatol.2019.1375>.
5. Haenssle HA, Fink C, Schneiderbauer R, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol*. 2018;29(8):1836–42. <https://doi.org/10.1093/annonc/mdy166>.
6. Brinker TJ, Hekler A, Enk AH, et al. Deep neural networks are superior to dermatologists in melanoma image classification. *Eur J Cancer*. 2019;119:11–7. <https://doi.org/10.1016/j.ejca.2019.05.023>.
7. Tschandl P, Rosendahl C, Akay BN, et al. Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. *JAMA Dermatol*. 2019;155(1):58–65. <https://doi.org/10.1001/jamadermatol.2018.4378>.
8. Haenssle HA, Fink C, Toberer F, et al. Man against machine reloaded: performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions in comparison with 96 dermatologists working under less artificial conditions. *Ann Oncol*. 2020;31(1):137–43. <https://doi.org/10.1016/j.annonc.2019.10.013>.
9. Information of International Skin Imaging Collaboration (ISIC). <https://challenge2019.isic-archive.com/>.

10. Hekler A, Utikal JS, Enk AH, et al. Superior skin cancer classification by the combination of human and artificial intelligence. *Eur J Cancer*. 2019;120: 114–21. <https://doi.org/10.1016/j.ejca.2019.07.019>.
11. Udrea A, Mitra GD, Costea D, et al. Accuracy of a smartphone application for triage of skin lesions based on machine learning algorithms. *J Eur Acad Dermatol Venereol*. 2020;34(3):648–55. <https://doi.org/10.1111/jdv.15935>.
12. Freeman K, Dinnis J, Chuchu N, et al. Algorithm based smartphone apps to assess risk of skin cancer in adults: systematic review of diagnostic accuracy studies. *BMJ*. 2020;368:m127. <https://doi.org/10.1136/bmj.m127>.
13. Hekler A, Utikal JS, Enk AH, et al. Pathologist-level classification of histopathological melanoma images with deep neural networks. *Eur J Cancer*. 2019;115: 79–83. <https://doi.org/10.1016/j.ejca.2019.04.021>.
14. George Y, Aldeen M, Garnavi R. Psoriasis image representation using patch-based dictionary learning for erythema severity scoring. *Comput Med Imaging Graph*. 2018;66:44–55. <https://doi.org/10.1016/j.compmedimag.2018.02.004>.
15. George Y, Aldeen M, Garnavi R. Automatic scale severity assessment method in psoriasis skin images using local descriptors. *IEEE J Biomed Health Inform*. 2020;24(2):577–85. <https://doi.org/10.1109/JBHI.2019.2910883>.
16. Fink C, Alt C, Uhlmann L, et al. Precision and reproducibility of automated computer-guided Psoriasis Area and Severity Index measurements in comparison with trained physicians. *Br J Dermatol*. 2019;180(2): 390–6. <https://doi.org/10.1111/bjd.17200>.
17. Meienberger N, Anzengruber F, Amruthalingam L, et al. Observer-independent assessment of psoriasis-affected area using machine learning. *J Eur Acad Dermatol Venereol*. 2020;34(6):1362–8. <https://doi.org/10.1111/jdv.16002>.
18. Maul LV, Meienberger N, Kaufmann L. Role of artificial intelligence in assessing the extent and progression of dermatoses. *Hautarzt*. 2020 Sep;71(9):677–85. <https://doi.org/10.1007/s00105-020-04657-5>.
19. Zhao S, Xie B, Li Y, et al. Smart identification of psoriasis by images using convolutional neural networks: a case study in China. *J Eur Acad Dermatol Venereol*. 2020;34(3):518–24. <https://doi.org/10.1111/jdv.15965>.
20. Gottlieb AB, Mease PJ, Kirkham B, et al. Secukinumab efficacy in psoriatic arthritis: machine learning and meta-analysis of four phase 3 trials. *J Clin Rheumatol*. 2020; <https://doi.org/10.1097/RHU.0000000000001302>.
21. Emam S, Du AX, Surmanowicz P, et al. Predicting the long-term outcomes of biologics in patients with psoriasis using machine learning. *Br J Dermatol*. 2020;182(5):1305–7. <https://doi.org/10.1111/bjd.18741>.
22. Tomalin LE, Kim J, Correa da Rosa J, et al. Early quantification of systemic inflammatory proteins predicts long-term treatment response to Tofacitinib and Etanercept. *J Invest Dermatol*. 2020;140(5): 1026–34. <https://doi.org/10.1016/j.jid.2019.09.023>.
23. De Guzman LC. Design and evaluation of a multi-model, multi-level artificial neural network for Eczema skin lesion detection. In: 3rd International conference on artificial intelligence, modelling and simulation (AIMS); 2015. p. 42–47. <https://doi.org/10.1109/AMIS.2015.17>.
24. Gustafson E, Pacheco J, Wehbe F, et al. A machine learning algorithm for identifying atopic dermatitis in adults from electronic health records. *IEEE Int Conf Healthc Inform*. 2017;2017:83–90. <https://doi.org/10.1109/ICHI.2017.31>.
25. Bobrova M, Taranki M, Kopanitsa G. Using neural networks for diagnosing in dermatology. *Stud Health Technol Inform*. 2019;261:211–6.
26. Han SS, Park I, Eun Chang S, et al. Augmented intelligence dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. *J Invest Dermatol*. 2020;140(9):1753–61. <https://doi.org/10.1016/j.jid.2020.01.019>.
27. Suter CNAAPM. Detection and quantification of hand eczema by visible spectrum skin pattern analysis. *Front Artif Intell Appl*. 2014;26(3):1101–2.
28. Melina A, Dinh NN, Tafuri B, et al. Artificial intelligence for the objective evaluation of acne investigator global assessment. *J Drugs Dermatol*. 2018;17(9): 1006–9.
29. Han SS, Park GH, Lim W, et al. Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: automatic construction of onychomycosis datasets by region-based convolutional deep neural network. *PLoS One*. 2018;13(1):e0191493. <https://doi.org/10.1371/journal.pone.0191493>.
30. Zang Q, Paris M, Lehmann DM, et al. Prediction of skin sensitization potency using machine learning approaches. *J Appl Toxicol*. 2017;37(7):792–805. <https://doi.org/10.1002/jat.3424>.
31. Liu M, Zhang J, Nie D, et al. Anatomical landmark based deep feature representation for MR images in brain disease diagnosis. *IEEE J Biomed Health Inform*. 2018;22(5):1476–85. <https://doi.org/10.1109/JBHI.2018.2791863>.
32. Seite S, Khammari A, Benzaquen M, et al. Development and accuracy of an artificial intelligence algorithm for acne grading from smartphone photographs. *Exp Dermatol*. 2019;28(11):1252–7. <https://doi.org/10.1111/exd.14022>.
33. Gomolin A, Netchiporouk E, Gniadecki R, et al. Artificial intelligence applications in dermatology: where do we stand? *Front Med (Lausanne)*. 2020;7:100. <https://doi.org/10.3389/fmed.2020.00100>.
34. Polesie S, Gillstedt M, Kittler H, et al. Attitudes towards artificial intelligence within dermatology: an international online survey. *Br J Dermatol*. 2020;183(1): 159–61. <https://doi.org/10.1111/bjd.18875>.

35. Winkler JK, Fink C, Toberer F, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatol.* 2019; <https://doi.org/10.1001/jamadermatol.2019.1735>.
36. Information of derma2go AG. www.darma2go.com/de/derma2go_de.
37. Information of E-derm-Consult GmbH. www.edermconsult.com.
38. Pangti R, Mathur J, Chouhan V, et al. A machine learning-based, decision support, mobile phone application for diagnosis of common dermatological diseases. *J Eur Acad Dermatol Venereol.* 2021;35(2): 536–45. <https://doi.org/10.1111/jdv.16967>.
39. Information of PASSION Dermatology. <https://www.telederm.ai/>.
40. Brunner PM, Guttman-Yassky E. Racial differences in atopic dermatitis. *Ann Allergy Asthma Immunol.* 2019;122(5):449–55. <https://doi.org/10.1016/j.anai.2018.11.015>.
41. Han SS, Kim MS, Lim W, et al. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *J Invest Dermatol.* 2018;138(7):1529–38. <https://doi.org/10.1016/j.jid.2018.01.028>.



Artificial Intelligence in Predicting Kidney Function and Acute Kidney Injury

40

Eiichiro Uchino, Noriaki Sato, and Yasushi Okuno

Contents

Introduction	562
AKI Overview	562
Background for AKI Prediction	562
ML Model for AKI Onset Prediction	563
Definition of AKI Event	563
Prediction Timepoint and Target Period	563
Input Features	563
ML Algorithms	572
Model Performance	572
The External Validity of the Models	572
Explainability of Models	573
Implementation Challenges	573
Conclusion	573
References	574

Abstract

Acute kidney injury (AKI) is a disease defined as an abrupt decline in kidney function and is a common complication in

hospitalized patients with high clinical significance. Recently, a model for predicting the onset of AKI clinical data by machine learning using electronic medical record data has attracted researchers' attention. The state-of-the-art model has achieved high discrimination performance of area under the curve ≥ 0.9 . Accordingly, these models are expected to be used for appropriate clinical intervention and disease prevention. In this chapter, we review the studies on AKI onset prediction and discuss their major issues. Since the event definitions, prediction timepoints, and prediction target periods used in the models

E. Uchino · N. Sato

Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto University, Kyoto, Japan

Department of Nephrology, Graduate School of Medicine, Kyoto University, Kyoto, Japan

Y. Okuno (✉)

Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto University, Kyoto, Japan

e-mail: okuno.yasushi.4c@kyoto-u.ac.jp

widely vary, we categorize them based on previous studies. In addition, we describe various input features; algorithms, including deep learning; model performance; and recent topics such as model explainability. Finally, we summarize achievements and challenges in implementing these models as clinical decision support tools.

Keywords

Kidney · Acute kidney injury (AKI) · Machine learning (ML) · Artificial intelligence (AI) · Deep learning (DL) · Electronic medical record (EMR) · Electronic health record (EHR) · Time series prediction · Disease onset prediction · Continuous prediction

Introduction

Kidneys play a central role in maintaining homeostasis in the human body. Their functions include the excretion of waste metabolites; regulation of fluid volume, electrolytes, and bone-mineral balance; and production of hematopoietic hormones. To measure overall kidney function, glomerular filtration rate (GFR) is accepted as the best indicator. In daily clinical practice, serum creatinine (SCr) levels are usually measured to estimate GFR. Using the SCr and estimated GFR values as indicators, diseases can be diagnosed. These include acute kidney injury (AKI), which is a rapid decrease in renal function, and chronic kidney disease (CKD), which is a chronic decrease in renal function. Predicting the onset and course of the disease is a clinically important challenge. Moreover, it is important to accurately predict the course of these clinical events and numerical values of their key indicators. The renal function and its temporal changes can be predicted numerically based on available clinical data. Thus, available models can be readily used to predict related transitions and events. Here, we review recent progress in these prediction models focusing on the onset of AKI as the most studied case.

AKI Overview

Acute kidney injury (AKI) is a common complication in hospitalized patients and is associated with high morbidity and mortality [1]. A recent meta-analysis has reported that one in five hospitalized patients develops AKI [2]. It is considered a major health issue, with 300,000 deaths annually in the United States, an average increase in hospital stays of 3.5 days, and a significant increase in medical expenses [3].

RIFLE (risk, injury, failure, loss of kidney function, and end-stage kidney disease) [4] and AKIN (Acute Kidney Injury Network) [5] criteria have been originally proposed and used for the diagnosis of AKI. More recently (in 2012), they have been replaced by KDIGO (Kidney Disease: Improving Global Outcomes) [6] criteria that are currently considered as a gold standard. According to the KDIGO criteria, AKI is diagnosed by meeting at least one of the following: (1) SCr increase by ≥ 0.3 mg/dL, (2) SCr increase to 1.5 times the baseline within 7 days, and (3) urine volume < 0.5 ml/kg/h for 6 h. AKI is also categorized into stages 1–3 by its severity [6].

Background for AKI Prediction

Early intervention of nephrologists, clinical specialists in this area, is crucial in AKI management, as discussed in the literature [7, 8]. More specifically, recent studies considered e-alert systems for automatic detection of AKI and accompanied intervention and demonstrated multiple clinical benefits, including shortening hospital stay [9]. On the other hand, it has been reported that simple, early detection alone does not reduce the mortality rate and subsequent renal function in randomized trials [10]. Thus, machine learning (ML) AKI models are considered a promising tool for early intervention and prevention.

Multivariate analysis has been widely used to assess risk factors of AKI. For instance, it has been efficient in several settings, including after cardiovascular surgery (patients with high risk of AKI [11–17]), after myocardial infarction [18], intensive care units (ICU) [19] sepsis patients in

ICU [20], pediatrics [21], pediatric ICU [22, 23], and general patients [24–27]. On the other hand, in recent years, there have been reports on machine learning models that comprehensively use available variables for more individual predictions with higher accuracy.

ML Model for AKI Onset Prediction

Multiple AKI onset prediction models for various situations have been reported using machine learning techniques. The appropriate parameters and performance of the models and their applications may vary depending on the situation. Common applications include general inpatients and ICU patients. Model parameters and performance may vary due to differences in pre-probability of AKI onset (patient condition severity) and in data acquisition intervals and content. Tables 1, 2, and 3 summarize previous studies on general inpatients, ICUs, and various other cases. Some important aspects of these models are discussed below.

Definition of AKI Event

As mentioned above, the KDIGO criteria are used as a standard for AKI diagnostics. Unfortunately, the urine volume data in general inpatients are difficult to collect and thus are rarely available. It results in a diagnosis of AKI based on the SCr values only in many studies. Some studies have adopted the AKIN classification or their own criteria. Also, there has been no consensus in definitions of baseline values for AKI diagnosis. Moreover, no studies have also examined the effect of different baseline definitions on the predictive performance of AKI models. Finally, the target AKI stage also differs depending on studies.

Prediction Timepoint and Target Period

When comparing existing studies, prediction timepoint (i.e., when to make the prediction or

the timepoint of instances to input into the model) and prediction target period are important. He J et al. [33] have examined how these conditions systematically affect model performance (details are given in Table 1). We have split the prediction timepoints into the following three categories: (I) case-control design, (II) at a certain point, and (III) continuous points. In addition, the prediction target period can be categorized into the following two categories: (i) within a certain window and (ii) by a certain event. The prediction framework of each model can be described by a combination of these categories (Fig. 1). (I) *Case-control design* is a retrospective design in which the onset of AKI is set as a positive label in advance, and then an instance of the negative label for the control is selected. Although this method can give a rough estimate of the contribution of features and prediction performance, it cannot be used in working models because the instance of the negative label cannot be set prospectively. (II) *At a certain point* is a standard design where data at a defined point, such as at the time of admission, is used as input to the model. In contrast, since the study of Koyner et al. [38], some researchers have used the design (III). This design continuously predicts multiple time series points for one patient, and these timepoints (e.g., on the day of admission, 1 day after, 2 days after) are collectively modeled. Since the models are trained and evaluated using all timepoints together, it is considered that the predictive ability and the contributing features are different from design (II). This continuous prediction design is specifically intended to predict the changes in a patient over time.

Input Features

Many models have used a large number of variables as input features, such as demographics, vital signs, laboratory results, medications, comorbidities, and admission and discharge records. The number of unique variables is very wide, ranging from about 100 to over 300,000. Since the structure of electronic medical record data differs depending on the system, it is difficult

Table 1 Reported models for general patients. In the algorithm column, the best performing algorithm is shown first. Abbreviations: AKI, acute kidney injury; BN, Bayesian network; CV, cross-validation; DT, decision tree; GBT, gradient boosted trees; k-NN, k-nearest neighbor; LASSO, least absolute shrinkage and selection operator;

References	Country	Setting	Patient's background	No. of patients	Prediction timepoint	Target period	Stage	AKI definition	Baseline SCR definition	Algorithm	Validation	Performance
Wu L et al. (2020) [28]	USA	Single-center	Adults, inpatients	76,957	(I) Case-control design, 24 h before AKI	(i) Within a certain window, 24 h	Any	KDIGO (without urine)	(1) Last value within 2 days before admission, (2) first value after admission	RF, LR, SVM, LogitBoost	10-fold CV	Age: 18–35, 0.784; 36–55, 0.766; 56–65, 0.754; >65, 0.725
Chen W et al. (2018) [29]	USA	Single-center	Adults, inpatients	76,957	(I) Case-control design, 24 h before AKI	(ii) By a certain event, by discharge	3	KDIGO (without urine)	(1) Last value within 2 days before admission, (2) first value after admission	RF, k-NN, DT, NN, ensemble of classifier	10-fold CV	Ensemble, 0.830; RF, 0.825
Cheng P et al. (2017) [30]	USA	Single-center	Adults aged 18–64 years, inpatients	33,703	(I) Case-control design, 1–5 days before AKI	(i) Within certain windows, 1–5 days	Any	KDIGO (without urine)	(1) Last value within 2 days before admission, (2) first value after admission	RF, AdaBoostM1, LR	10-fold CV	1-day model: RF, 0.765; AdaBoostM1, 0.763; LR, 0.751
Chen YS et al. (2020) [31]	China	Single-center	Adults, inpatients	7930	(I) Case-control design, 1–5 days before AKI	(i) Within certain windows, 1–5 days	Any	AKIN (without urine)	N/A	k-NN	Split training (80%), test (20%) sets	0.866
Wu L et al. (2018) [32]	USA	Single-center	Adults, inpatients	76,957	Before AKI (details unknown)	(ii) By a certain event, by discharge	Any, ≥2, 3	KDIGO (without urine)	(1) Last value within 2 days before admission, (2) first value after admission	RF	10-fold CV	Any stage, 0.76; stage ≥2, 0.80; stage 3, 0.82
He J et al. (2019) [33]	USA	Single-center	Adults, inpatients	76,957	(I) Case-control design, 24 h before AKI (model 1); (II) at certain points, at 1, 2, 3, 7,	(i) Within certain windows, 24 h (models 1 and 4); (II) at certain points, at 1, 2, 3, 7,	Any	KDIGO (without urine)	(1) Last value within 2 days before admission, (2) first value after admission	RF + LR ensemble, RF, LR, NB, BN	10-fold CV	Model 1, 0.744; model 2, 0.734; model 3 within 1, 2, 7, and 30 days, 0.764, 0.727, 0.722, and 0.734;

			admission (models 2 and 3); (ii) by a certain event, discharge (model 2)	15, and 30 days (model 3); (ii) by a certain event, discharge (model 2)	KDIGO (without urine)	XGB, LASSO	Temporal validation	model 4 at days 1 and 4, 0.679 and 0.600
Hsu CN et al. (2020) [34]	Taiwan	Multi-center	Adults, inpatients	234,867	(II) At a certain point, at admission	Any, ≥ 2 (community acquired)	(1) Last value within 7 days, (2) mean value within 8–90 days	Any stage, 0.761; stage ≥ 2 , 0.818
Kate RJ et al. (2016) [35]	USA	Multi-center (15 sites)	Adults aged ≥ 60 years, inpatients	25,521	(II) At a certain point, 24 h after admission	Any	AKIN	Last value within 48 h
Cronin RM et al. (2015) [36]	USA	Multi-center (116 Department of Veterans Affairs hospitals)	Adults, inpatients	1,620,898	(II) At a certain point, at 48 h after admission	Any, ≥ 2	KDIGO (without urine)	Mean value within 7–365 days
Davis SE et al. (2017) [37]	USA	Multi-center (all Department of Veterans Affairs hospitals)	Inpatients	1,841,951	(II) At a certain point, at 48 h after admission	Any	KDIGO (without urine)	N/A
Koyner JL et al. (2016) [38]	USA	Multi-center (5 sites)	Adults, inpatients	202,961	(III) Continuous points, every 12 h	Any, ≥ 2 , 3	KDIGO (without urine)	First value on admission, and then updated on a rolling basis by KDIGO criteria
Koyner JL et al. (2018) [39]	USA	Single-center	Adults, inpatients	121,158	(III) Continuous points, every measurement 24 and 48 h	Any	KDIGO (without urine)	First value on admission, and then updated on a rolling basis of 48 h or 7 days

(continued)

Table 1 (continued)

References	Country	Setting	Patient's background	No. of patients	Prediction timepoint	Target period	Stage	AKI definition	Baseline SCr definition	Algorithm	Validation	Performance
Mohamadou H et al. (2018) [40]	USA	Multi-center (2 sites; models were independently constructed and evaluated)	Adults, inpatients	68,319	(III) Continuous points, at every test	(i) Within certain windows; 12, 24, 48 and 72 h	≥ 2	National Health Service England AKI algorithm and KDIGO	(1) Minimum value within 7 days of the index hospitalization, (2) median value within 8–365 days	XGB 3-fold CV	12-h model, 0.80; 24-h, 0.79; 48-h, 0.76; 72-h, 0.73	
Simonov M et al. (2019) [41]	USA	Multi-center (3 sites)	Adults, inpatients	169,859	(III) Continuous points, at every record	(i) Within a certain window, 24 h	Any	KDIGO (without urine)	Lowest value within 7 days	LR	External validation at other sites	0.74
Tomašev N et al. (2019) [42]	USA	Multi-center (172 admission sites and 1062 outpatient sites)	Adults, inpatients	703,782	(III) Continuous points, every 6 h	(i) Within certain windows; 24, 48, and 72 h	Any, ≥ 2 , 3	KDIGO (without urine)	(1) Median value within 1 year, (2) estimation by MDRD formula	Multi-task RNN, GBT, RF, LR, etc.	Split training (80%), validation (5%), calibration (5%), test (10%) sets	Any stage, 0.92; stage ≥ 2 , 0.95; stage 3, 0.980
Song X, Waitman LR et al. (2020) [43]	USA	Single-center	Adults, inpatients	76,957	(III) Continuous points, for every encounter	(i) Within a certain window, 24 h	Any	KDIGO (without urine)	(1) Last value within 2 days before admission, (2) first value after admission	RF	Split training (80%), test (20%) sets	Benchmark model, 0.81; in anonymized dataset, 0.78
Song X, Alan SL et al. (2020) [44]	USA	Multi-center (20 systems)	Adults, inpatients	153,821	(III) Continuous points, days 1–7 after admission	(i) Within certain windows; 24 and 48 h	Any, ≥ 2 , 3	KDIGO (without urine)	Latest value	GBT, LASSO	Temporal validation	48-h model, any stage, 0.76, 0.81; stage ≥ 2 , 0.81; stage 3, 0.87

Table 2 Reported models for intensive care unit patients. In the algorithm column, the best performing algorithm is shown first. Abbreviations: CNN, convolutional neural network; DNN, deep neural network; MDRD, the Modification of Diet in Renal Disease; MLP, multiple-layer perceptron. Other abbreviations are the same as in Table 1

References	Country	Setting	Patient's background	No. of patients	Prediction timepoint	Target period	AKI definition	Baseline SCr definition	Algorithm	Validation	Performance
Li Y et al. (2018) [45]	USA	Single-center (MIMIC-III)	Adults, ICU patients	14,470	(II) At a certain point, at admission	(i) Within a certain window, 24 h	Any KDIGO (detail unknown about urine)	N/A	CNN, LR, RF, GBT	Split training (70%), test (30%) sets	0.779
Flechet M et al. (2019) [46]	Belgium	Multi-center (5 sites)	Adults, ICU patients	252	(II) At certain points, at first morning and 24 h after admission	(ii) By a certain event, 1 week after ICU admission	≥2 KDIGO (with urine only in development cohort)	(1) Lowest value within 3 months not including admission. (2) estimation by MDRD formula	RF	External validation in another site	At first morning, 0.75; at 24 h, 0.89
Zimmerman LP et al. (2019) [47]	USA	Single-center (MIMIC-III)	Adults, ICU patients	23,950	(II) At a certain point, at 24 h after admission	(i) Within a certain window, 48 h	Any KDIGO	Lowest value at day 1 of admission	MLP, LR, RF	10-time fivefold CV	MLP, 0.796; LR, 0.783; RF, 0.779
Sun M et al. (2019) [48]	USA	Single-center (MIMIC-III)	Adults, ICU patients	14,469	(II) At a certain point, at 24 h after admission	(i) Within a certain window, 48 h	Any KDIGO	Lowest value at day 1 of admission	SVM, CNN, LR, RF, NB	Split training (70%), test (30%) sets	0.835
Gong K et al. (2021) [49]	USA	Single-center (MIMIC-III)	Adults, ICU patients	46,520	(II) At certain points, at 24 h after admission	(i) Within a certain window, 48 h	Any KDIGO (without urine)		XGB, ensemble of LR and RF	Split training (80%), test (20%) sets	0.774
Xu Z et al. (2019) [50]	USA	Single-center (MIMIC-III)	Adults, ICU patients	58,976	(II) At certain points; at 24, 48, 72, 96, 120, and 144 h after admission	(i) Within a certain window, 7 days	Any KDIGO	N/A	XGB, LR, L2 regulated LR, RF	5-fold CV	At 24 h, 0.75
Morid MA et al. (2020) [51]	USA	Single-center (MIMIC-III)	Adults, ICU patients	22,542	(II) At a certain point, at 48 h after admission	(ii) By a certain event, by discharge	Any AKIN (without urine)	Lowest value after admission	RF, XGB, BN, SVM, LR, NB, k-NN, NN	20-fold CV	With temporal features, 0.809; only last recorded value, 0.589

(continued)

Table 2 (continued)

References	Country	Setting	Patient's background	No. of patients	Prediction timepoint	Target period	AKI Stage definition	Baseline SCr definition	Algorithm	Validation	Performance
Parreco J et al. (2019) [52]	USA	Multi-center (eICU Collaborative Research Database)	Adults, ICU patients	151,098	(II) At a certain point, at 3 days after admission	(ii) By a certain event, by discharge	Any KDIGO	N/A	GBT, LR, DNN	10-fold CV	GBT, 0.834; LR, 0.827; DNN, 0.817
Chiofolo C et al. (2019) [53]	USA	Single-center	Adults, ICU patients	6530	(III) continuous points, every 15 min	(i) Within a certain window, 6 h	Any AKIN ≥ 2	Defined as the median value of creatinine during 180 days prior to the index ICU admission	RF	Split training (70%), test (30%) sets	Any stage, 0.882; stage ≥ 2 , 0.878
Wang Y et al. (2020) [54]	China, USA	Multi-center (ICUC and MIMIC-III)	Adults, ICU patients	65,205	(III) Continuous points, every day	(i) Within certain windows; 24 and 48 h	Any KDIGO (without urine)	N/A	Originally proposed Ensemble Time Series Model (ETSM), XGB, NB, k-NN, AdaBoost, RF	10-time split training (60%), test (40%) sets	ICUC, 24-h model, 0.81; 48-h, 0.78; MIMIC III, 24-h, 0.95; 48-h, 0.95

Table 3 Reported models for various other situations. In the algorithm column, the best performing algorithm is shown first. GA, genetic algorithm; GAM, generalized additive model. Other abbreviations are the same as in Tables 1 and 2

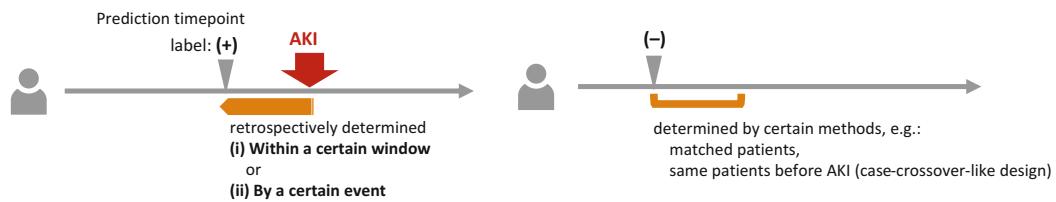
References	Country	Setting	Patient's background	No. of patients	Prediction timepoint	Target period	Stage	AKI definition	Baseline SCR	Algorithm	Validation	Performance
Ibrahim NE et al. (2019) [55]	USA	Single-center	Patients undergoing coronary angiography	889	(II) At a certain point, at contrast exposure	(i) Within a certain window, 7 days	Any	KDIGO (without urine)	N/A	L1-regularized LR	400-time split training (80%), test (20%) sets	0.79
Huang C et al. (2018) [56]	USA	Multi-center (≥ 1500 sites in NCDR CathPCI Registry)	Adult, patients undergoing PCI procedures	1,917,960	(II) At a certain point, before procedure	(ii) By a certain event, by discharge	Any	AKIN	Latest value before procedure	XGB, LR	Temporal validation	0.759
Huang C et al. (2019) [57]	USA	Multi-center (≥ 1694 sites in NCDR CathPCI Registry)	Adult, patients undergoing PCI procedures	3,038,557	(II) At a certain point, before procedure	(ii) By a certain event, by discharge	Any	Originally defined; SCR increase by 0.3, 0.5, and 1.0 mg/dL	Latest value before procedure	GAM	Temporal validation	0.3 mg/dL model, 0.777
Tseng PY et al. (2020) [58]	Taiwan	Single-center	Inpatients undergoing cardiac surgery	671	(II) At a certain point, at 4 h after surgery	(i) Within a certain window, 7 days	Any	KDIGO (without urine)	Latest value before surgery	XGB + RF ensemble, XGB, RF, SVM, LR	Split training (70%), test (30%) sets	Ensemble, 0.843; XGB, 0.837; RF, 0.839; LR, 0.806
Rank N et al. (2020) [59]	Germany	Single-center	Adults, inpatients undergoing cardiothoracic surgery	2,572	(III) Continuous points, every 15 min after surgery	(ii) By a certain event, 7 days after surgery	≥ 2	KDIGO	Latest value before surgery	RNN	Split training (85%), test (15%) sets	0.893
Thottakkara P et al. (2016) [60]	USA	Single-center	Adults, inpatients undergoing major surgery	50,318	(II) At a certain point, at surgery	(i) Within a certain window, 7 days	Any	KDIGO (without urine)	N/A	GAM, SVM, LR, NB	50-time split training (70%), test (30%) sets	GAM, 0.858; SVM, 0.857; LR, 0.853; NB, 0.819
Adhikari L et al. (2019) [61]	USA	Single-center	Adult, patients undergoing surgery	2911	(II) At certain points, before and after surgery	(ii) By certain events; 3 and 7 days after	Any	KDIGO (detail unknown about urine)	(1) Minimum value within 7 days of the index hospitalization, (2) median	RF	Split training (70%), test (30%) sets	Preoperative 7-day model, 0.84; postoperative 7-day, 0.86

(continued)

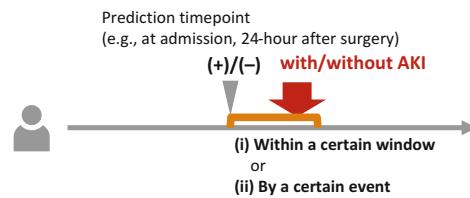
Table 3 (continued)

References	Country	Setting	Patient's background	No. of patients	Prediction timepoint	Target period	Stage	AKI definition	Baseline SCr definition	Algorithm	Validation	Performance
Lei VJ et al. (2019) [62]	USA	Multi-center (4 sites)	Adult, patients undergoing major noncardiac surgery	42,615	(II) At certain points; at prehospitalization, before and after surgery	Any	KDIGO	Lowest value within 1 year	GBT, LR with Elastic Net selection, RF	Split training (60%), validation (20%), test (20%) sets	With prehospitalization variables, 0.712; + preoperative variables, 0.804; + intraoperative variables, 0.817	
Jeon N et al. (2019) [63]	USA	Multi-center (2 sites)	Adults, inpatients with nephrotoxic medication	62,561	(II) At certain points, at 1–5 days after admission	≥2	KDIGO (without urine)	Latest value within 2 days	LR	100 bootstrap resamples	At day 1–5 models: 0.783, 0.808, 0.798, 0.779, 0.779	
Martinez DA et al. (2020) [64]	USA	Multi-center (3 sites)	Adults, emergency department inpatients	59,792	(II) At a certain point, at admission	Any, ≥2	KDIGO (without urine)	Value at emergency department arrival	RF	10-time Monte Carlo CV	Any stage, 24-h model, 0.80; 48-h, 0.76; 72-h, 0.74; stage ≥2, 24-h, 0.81; 48-h, 0.77; 72-h, 0.75	
Weisenthal SJ et al. (2018) [65]	USA	Single-center	Adult, rehospitalized patients	34,505	(II) At a certain point, at hospital re-entry	Any	ICD9 codes or KDIGO (without urine)	First value on admission and then updated on a rolling basis by KDIGO criteria	GBT, penalized LR, RNN	50-time fivefold CV	0.867	
Park N et al. (2018) [66]	Korea	Single-center	Cancer inpatients	21,022	(III) Continuous points, every SCr measurement	Within a certain window, 14 days	Max SCr values (regression) into any stage	Modified KDIGO (without urine), longer window of 3 weeks	RF, linear regression	Nested 3 × threefold CV	AUC, N/A; precision, 78.9%; recall, 75.1%; F-measure, 75.8%	
Sandokaji I et al. (2020) [67]	USA	Multi-center (2 sites)	Child, inpatients	8473	(III) Continuous points, every test	Any, ≥2	KDIGO (without urine) with absolute SCr >0.5 mg/dl	Lowest value within 48 h or 7 days	GA	Internally splitting training (70%). Test (30%) sets; externally in the other site	Any stage, 0.76; stage ≥2, 0.79	

(I) Case-control design



(II) At a certain point



(III) Continuous points

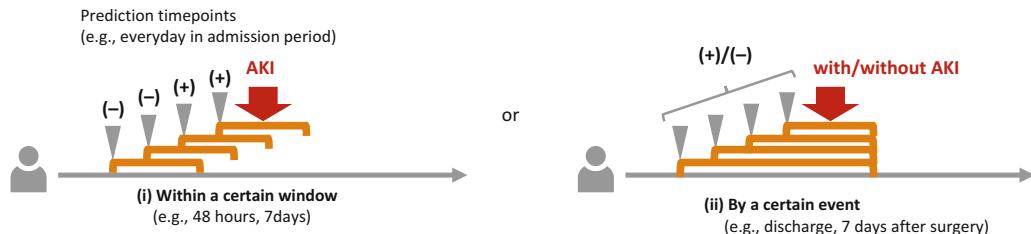


Fig. 1 Prediction timepoint and target period. Problem settings can be categorized using the combinations of prediction timepoint (arrowhead), (I) case-control design,

(II) at a certain point, and (III) continuous points, and target period (orange), (i) within a certain window and (ii) by a certain event

to unify the obtained feature set. Thus, it is practically necessary to make a feature set each time depending on the location where the model is built and operated. For example, in the postoperative AKI model, it is important to incorporate appropriate variables according to the settings, such as intraoperative time series features [61, 62]. As a factor analysis in an experimental setting, there has been a report in which proteomics data have been added as omics data to electronic medical record data [55]. When constructing the model, a feature selection process is sometimes required. Wu et al. [32] have compared several common feature selection methods and found that the performance of the model did not change significantly, but the importance ranking fluctuated significantly.

In most cases, electronic medical record data has irregular measurement intervals. Therefore,

when inputting a model, the representative values in some windows are often acquired in units of several hours to several days. If necessary, the interpolation with the feedforward method or median values is performed for missing data. Some models [42] of recurrent neural networks (RNNs) handle time series data in its original format. In other models, the derived features (mean, variance, maximum/minimum, or trends) are selectively or comprehensively calculated within a window and are used as input data. For instance, some researchers considered time series trend detection that improved the model performance [51] and a model taking into account time series changes in drug combinations [54]. In addition, recent studies (technical verification stage) have succeeded in creating features from unstructured data from clinical notes using natural language processing (NLP) [45]. Other researchers

attempted to verify performance changes using similar NLP-derived features added to the conventional feature set [48].

ML Algorithms

The most commonly used algorithms are tree-based ones, such as random forest (RF) and gradient boosted trees (GBT). These algorithms generally demonstrate high performance. Other common methods include logistic regression (LR), support vector machine (SVM), k-nearest neighbor (k-NN), naïve Bayes (NB), and deep neural networks (DNN). Tomašev et al. [42] have achieved the highest performance of all studies using multi-tasking RNNs. They have also provided the most comprehensive performance comparison of each method and concluded that their method demonstrated the best performance. On the other hand, in many studies, including later ones, other algorithms such as extreme gradient boosting (XGB) have shown the best performance, and the optimal method is still unclear. One of the factors may be that the optimal design of DNN is a complex process compared to other general machine learning models. In addition, the sample size for each study and the characteristics of the dataset may also influence method performance.

Model Performance

The area under the curve (AUC) of the receiver operating characteristic (ROC) curve is used as a common evaluation index in most studies. AUC value ranges from about 0.7 to 0.9, as shown in Tables 1, 2, and 3. Many studies show that the AUC value tends to be higher as the target AKI stage is higher or the prediction target period is shorter. Thus, for the model robustness, there has been a statement that it is desirable to predict AKI of stage ≥ 2 [68]. Taking the 48-h model as an example, the best performance was achieved by Tomašev N et al. in [42] and amounted to 0.921 for any stage AKI and 0.957 for stage ≥ 2 . Some

studies have also evaluated the precision, recall (sensitivity), or specificity of the model for some thresholds. In particular, Tomašev N et al. [42] have argued that a setting with a precision of 33% or 25% is realistic for practical applications. For example, in their 48-h model, about 55% of AKI patients could be picked up under the condition of precision 33%, i.e., there are two false-positive predictions for a true positive prediction.

The prediction performance is determined by multiple factors. In particular, precision value is particularly affected by prior probabilities. For example, although it has been reported that the difference in AUC value between general ward and ICU is small [39] since ICU patients have a higher prior probability of AKI, its effect on the precision and thresholds should still be considered. In another example, focusing on the age of patients, Wu et al. [28] have reported that the older they are, the lower the predictive performance and the greater the variation in the top factors contributing to prediction accuracy.

The External Validity of the Models

Few studies have used the models validated with external datasets. For example, the model by Koyner et al. [39] has been validated at multiple institutions, and stable results have been obtained [69]. Song et al. [44] have reported a model that predicts the transportability of AI between different locations and validated it for the AKI prediction. However, the stability of the external setting in many models is still unclear. MIMIC-III (Medical Information Mart for Intensive Care III) [70] data in ICU patients is the only common dataset that can be widely used, making it difficult to compare performance between studies. This issue is especially pressing when discussing model performance in general patients, for which no open data is available. Sharing anonymous data should help, but there are many barriers to its practical applications. Song et al. [43] have examined the model performance using the anonymous data for AKI prediction and found that the performance was maintained at a slightly reduced

level. However, due to the sparse nature of medical data, the risk of re-identification is still high.

Many studies have used years of data to train and validate models. As these are time-related data, it is an important issue whether the models can fit future data. For evaluation of the reported models, various methods are used. For instance, temporal validation splits the dataset by time, and the simple cross-validation method does not consider time at all. To address this issue, Davis et al. [37] have analyzed discrimination performance in a 9-year validation cohort. The AUC values were generally maintained in chronological order, but the magnitude of overprediction increased over time in regression models. Thus, the authors emphasized the importance of protocols for updating the models.

Explainability of Models

Multiple studies [38, 39, 41, 42, 50, 52, 63, 65] have discussed the explainability of the models. Feature importance values calculated from tree-based algorithms or permutation importance values have been used in the standard methods. SCr and its derivatives, including mean, change, and slope, were the most frequently used features. A wide variety of other variables are listed as top features, e.g., blood urea nitrogen, hemoglobin, bicarbonate, chloride, history of malignant neoplasm of the kidney, emergency department visits, and vancomycin use.

Recently there have been some studies utilizing Shapley additive explanations (SHAP) [71] for feature analysis. For instance, Song et al. [44] have illustrated partial dependence of top features by the method, and Tseng et al. [58] have shown that SHAP picked different top time series-related features compared to the conventional methods. In addition to those top global features mentioned, Gong et al. [49] have also provided an example of interpretation at the individual level using SHAP. However, the clinical interpretation of the model output (e.g., in which cases high output values are predicted and obtained) is

still insufficiently understood, thus hindering its practical applications.

Implementation Challenges

The prediction results provided by high-performance models are likely to generate unprecedented clinical processes. Interestingly, Rank et al. [59] have analyzed the performance of AKI prediction by clinicians. Even experienced clinicians showed an AUC of 0.745, which was lower than the result of many predictive models. Sutherland et al. [68] have illustrated “renal dashboard” as a screen sample to provide AKI risk scores and information about the potential intervention. However, only a few studies have implemented the prediction models as clinical tools. For instance, Ugwuowo et al. [72] have implemented and verified a model from [41] in a single-center electronic medical record system for 1.5 years. They have reported that 2856 patients at risk of 15% were extracted, and 18.9% of them actually developed AKI within 24 h. However, the impact of these models on the prevention of AKI has not been evaluated. In pediatric patients, Driest et al. [73] have conducted a randomized trial of the operation of another model [21]. The alerts increased the number of SCr tests in the ICU but did not change the incidence or severity of AKI. In addition, there was no difference in those indicators of patients in the ward. Future studies should consider the implementation of these models and their effects on the behavior and performance of clinicians, as well as the outcomes, including AKI onset prevention, renal function, and survival rate.

Conclusion

In this chapter, we have reviewed previous studies on AKI onset prediction and discussed their key issues. In particular, we have summarized technical issues, including algorithms, feature engineering, and explanations. We have demonstrated that further development of these technical elements

and their applications is required. Finally, we have emphasized the importance of implementation issues for future studies.

References

- Li PKT, Burdmann EA, Mehta RL. Acute kidney injury: global health alert. *Kidney Int.* 2013;83:372–6. <https://doi.org/10.1038/ki.2012.427>.
- Susantitaphong P, Cruz DN, Cerdá J, Abulfaraj M, Alqahtani F, Koulouridis I, et al. World incidence of AKI: a meta-analysis. *Clin J Am Soc Nephrol.* 2013;8:1482–93. <https://doi.org/10.2215/CJN.00710113>.
- Lewington AJP, Cerdá J, Mehta RL. Raising awareness of acute kidney injury: a global perspective of a silent killer. *Kidney Int.* 2013;84:457–67. <https://doi.org/10.1038/ki.2013.153>.
- Bellomo R, Ronco C, Kellum JA, Mehta RL, Palevsky P. Acute renal failure – definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group. *Crit Care.* 2004;8:R204. <https://doi.org/10.1186/cc2872>.
- Mehta RL, Kellum JA, Shah SV, Molitoris BA, Ronco C, Warnock DG, et al. Acute kidney injury network: report of an initiative to improve outcomes in acute kidney injury. *Crit Care.* 2007;11:R31. <https://doi.org/10.1186/cc5713>.
- KDIGO. KDIGO clinical practice guideline for acute kidney injury. *Kidney Int Suppl.* 2012;2:1–138.
- Mehta RL, McDonald B, Gabbai F, Pahl M, Farkas A, Pascual MTA, et al. Nephrology consultation in acute renal failure does timing matter? *Am J Med.* 2002;113:456–61. [https://doi.org/10.1016/s0002-9343\(02\)01230-5](https://doi.org/10.1016/s0002-9343(02)01230-5).
- Endre ZH. The role of nephrologist in the intensive care unit. *Blood Purif.* 2017;43:78–81. <https://doi.org/10.1159/000452318>.
- Selby NM, Casula A, Lamming L, Stoves J, Samarasinghe Y, Lewington AJ, et al. An organizational-level program of intervention for AKI: a pragmatic stepped wedge cluster randomized trial. *J Am Soc Nephrol.* 2019;30:505–15. <https://doi.org/10.1681/asn.2018090886>.
- Wilson PF, Shashaty M, Testani J, Aqeel I, Borovskiy Y, Ellenberg SS, et al. Automated, electronic alerts for acute kidney injury: a single-blind, parallel-group, randomised controlled trial. *Lancet.* 2015;385:1966–74. [https://doi.org/10.1016/S0140-6736\(15\)60266-5](https://doi.org/10.1016/S0140-6736(15)60266-5).
- Thakar CV, Arrigain S, Worley S, Yared J-P, Paganini EP. A clinical score to predict acute renal failure after cardiac surgery. *J Am Soc Nephrol.* 2005;16:162–8. <https://doi.org/10.1681/asn.2004040331>.
- Mehta RH, Grab JD, O'Brien SM, Bridges CR, Gammie JS, Haan CK, et al. Bedside tool for predicting the risk of postoperative dialysis in patients undergoing cardiac surgery. *Circulation.* 2006;114:2208–16. <https://doi.org/10.1161/circulationaha.106.635573>.
- Aronson S, Fontes ML, Miao Y, Mangano DT. Risk index for perioperative renal dysfunction/failure. *Circulation.* 2007;115:733–42. <https://doi.org/10.1161/circulationaha.106.623538>.
- Brown JR, Cochran RP, Leavitt BJ, Dacey LJ, Ross CS, MacKenzie TA, et al. Multivariable prediction of renal insufficiency developing after cardiac surgery. *Circulation.* 2007;116:I-139–43. <https://doi.org/10.1161/circulationaha.106.677070>.
- Wijeysundera DN, Karkouti K, Dupuis J-Y, Rao V, Chan CT, Granton JT, et al. Derivation and validation of a simplified predictive index for renal replacement therapy after cardiac surgery. *JAMA.* 2007;297:1801–9. <https://doi.org/10.1001/jama.297.16.1801>.
- Palomba H, de Castro I, Neto ALC, Lage S, Yu L. Acute kidney injury prediction following elective cardiac surgery: AKICS score. *Kidney Int.* 2007;72:624–31. <https://doi.org/10.1038/sj.ki.5002419>.
- Simonini M, Lanzani C, Bignami E, Casamassima N, Frati E, Meroni R, et al. A new clinical multivariable model that predicts postoperative acute kidney injury: impact of endogenous ouabain. *Nephrol Dial Transplant.* 2014;29:1696–701. <https://doi.org/10.1093/ndt/gfu200>.
- Zambetti BR, Thomas F, Hwang I, Brown AC, Chumpia M, Ellis RT, et al. A web-based tool to predict acute kidney injury in patients with ST-elevation myocardial infarction: development, internal validation and comparison. *PLoS One.* 2017;12:e0181658. <https://doi.org/10.1371/journal.pone.0181658>.
- Malhotra R, Kashani KB, Macedo E, Kim J, Bouchard J, Wynn S, et al. A risk prediction score for acute kidney injury in the intensive care unit. *Nephrol Dial Transplant.* 2017;32:814–22. <https://doi.org/10.1093/ndt/gfx026>.
- Deng F, Peng M, Li J, Chen Y, Zhang B, Zhao S. Nomogram to predict the risk of septic acute kidney injury in the first 24 h of admission: an analysis of intensive care unit data. *Ren Fail.* 2020;42:428–36. <https://doi.org/10.1080/0886022x.2020.1761832>.
- Wang L, McGregor TL, Jones DP, Bridges BC, Fleming GM, Shirey-Rice J, et al. Electronic health record-based predictive models for acute kidney injury screening in pediatric inpatients. *Pediatr Res.* 2017;82:465–73. <https://doi.org/10.1038/pr.2017.116>.
- Sanchez-Pinto LN, Khemani RG. Development of a prediction model of early acute kidney injury in critically ill children using electronic health record data. *Pediatr Crit Care Med.* 2016;17:508–15. <https://doi.org/10.1097/pcc.0000000000000750>.
- Raman S, Tai CW, Marsney RL, Schibler A, Gibbons K, Schlapbach LJ. Prediction of acute kidney injury on admission to pediatric intensive care. *Pediatr Crit Care Med.* 2020;21:811–9. <https://doi.org/10.1097/pcc.0000000000002411>.
- Drawz PE, Miller RT, Sehgal AR. Predicting hospital-acquired acute kidney injury – a case-controlled study. *Ren Fail.* 2009;30:848–55. <https://doi.org/10.1080/08860220802356515>.

25. Matheny ME, Miller RA, Ikizler TA, Waitman LR, Denny JC, Schildcrout JS, et al. Development of inpatient risk stratification models of acute kidney injury for use in electronic health records. *Med Decis Mak.* 2010;30:639–50. <https://doi.org/10.1177/0272989X10364246>.
26. Forni LG, Dawes T, Sinclair H, Cheek E, Bewick V, Dennis M, et al. Identifying the patient at risk of acute kidney injury: a predictive scoring system for the development of acute kidney injury in acute medical patients. *Nephron Clin Pract.* 2013;123:143–50. <https://doi.org/10.1159/000351509>.
27. Bedford M, Stevens P, Coulton S, Billings J, Farr M, Wheeler T, et al. Development of risk models for the prediction of new or worsening acute kidney injury on or during hospital admission: a cohort and nested study, vol. 4. *Health Services and Delivery Research;* 2016. p. 1–160. <https://doi.org/10.3310/hsdr04060>.
28. Wu L, Hu Y, Zhang X, Chen W, Yu ASL, Kellum JA, et al. Changing relative risk of clinical factors for hospital-acquired acute kidney injury across age groups: a retrospective cohort study. *BMC Nephrol.* 2020;21:321. <https://doi.org/10.1186/s12882-020-01980-w>.
29. Chen W, Hu Y, Zhang X, Wu L, Liu K, He J, et al. Causal risk factor discovery for severe acute kidney injury using electronic health records. *BMC Med Inform Decis Mak.* 2018;18:13. <https://doi.org/10.1186/s12911-018-0597-7>.
30. Cheng P, Waitman LR, Hu Y, Liu M. Predicting inpatient acute kidney injury over different time horizons: how early and accurate? *AMIA Annu Symp Proc.* 2017;2017:565–74.
31. Chen Y-S, Chou C-Y, Chen ALP. Early prediction of acquiring acute kidney injury for older inpatients using most effective laboratory test results. *BMC Med Inform Decis Mak.* 2020;20:36. <https://doi.org/10.1186/s12911-020-1050-2>.
32. Wu L, Hu Y, Liu X, Zhang X, Chen W, Yu ASL, et al. Feature ranking in predictive models for hospital-acquired acute kidney injury. *Sci Rep.* 2018;8:17298. <https://doi.org/10.1038/s41598-018-35487-0>.
33. He J, Hu Y, Zhang X, Wu L, Waitman LR, Liu M. Multi-perspective predictive modeling for acute kidney injury in general hospital populations using electronic medical records. *JAMIA Open.* 2019;2: ooy043. <https://doi.org/10.1093/jamiaopen/ooy043>.
34. Hsu C-N, Liu C-L, Tain Y-L, Kuo C-Y, Lin Y-C. Machine learning model for risk prediction of community-acquired acute kidney injury hospitalization from electronic health records: development and validation study. *J Med Internet Res.* 2020;22:e16903. <https://doi.org/10.2196/16903>.
35. Kate RJ, Perez RM, Mazumdar D, Pasupathy KS, Nilakantan V. Prediction and detection models for acute kidney injury in hospitalized older adults. *BMC Med Inform Decis Mak.* 2016;16:39. <https://doi.org/10.1186/s12911-016-0277-4>.
36. Cronin RM, VanHouten JP, Siew ED, Eden SK, Fihn SD, Nielson CD, et al. National Veterans Health Administration inpatient risk stratification models for hospital-acquired acute kidney injury. *J Am Med Inform Assoc.* 2015;22:1054–71. <https://doi.org/10.1093/jamia/ocv051>.
37. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Inform Assoc.* 2017;24:1052–61. <https://doi.org/10.1093/jamia/ocx030>.
38. Koyner JL, Adhikari R, Edelson DP, Churpek MM. Development of a multicenter ward-based AKI prediction model. *Clin J Am Soc Nephrol.* 2016;11: 1935–43. <https://doi.org/10.2215/CJN.00280116>.
39. Koyner JL, Carey KA, Edelson DP, Churpek MM. The development of a machine learning inpatient acute kidney injury prediction model. *Crit Care Med.* 2018;46:1070–7. <https://doi.org/10.1097/CCM.0000000000003123>.
40. Mohamadlou H, Lynn-Palevsky A, Barton C, Chettipally U, Shieh L, Calvert J, et al. Prediction of acute kidney injury with a machine learning algorithm using electronic health record data. *Can J Kid Heal Dis.* 2018;5:2054358118776326. <https://doi.org/10.1177/2054358118776326>.
41. Simonov M, Ugwuowo U, Moreira E, Yamamoto Y, Biswas A, Martin M, et al. A simple real-time model for predicting acute kidney injury in hospitalized patients in the US: a descriptive modeling study. *PLoS Med.* 2019;16:e1002861. <https://doi.org/10.1371/journal.pmed.1002861>.
42. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature.* 2019;572:116–9. <https://doi.org/10.1038/s41586-019-1390-1>.
43. Song X, Waitman LR, Hu Y, Luo B, Li F, Liu M. The impact of medical big data anonymization on early acute kidney injury risk prediction. *AMIA Jt Summits Transl Sci Proc.* 2020;2020:617–25.
44. Song X, Yu ASL, Kellum JA, Waitman LR, Matheny ME, Simpson SQ, et al. Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. *Nat Commun.* 2020;11:5668. <https://doi.org/10.1038/s41467-020-19551-w>.
45. Li Y, Yao L, Mao C, Srivastava A, Jiang X, Luo Y. Early prediction of acute kidney injury in critical care setting using clinical notes. *IEEE Int Conf Bioinformatics Biomed.* 2018;2018:683–6. <https://doi.org/10.1109/bibm.2018.8621574>.
46. Flechet M, Falini S, Bonetti C, Güiza F, Schetz M, den Berghe GV, et al. Machine learning versus physicians' prediction of acute kidney injury in critically ill adults: a prospective evaluation of the AKI predictor. *Crit Care.* 2019;23:282. <https://doi.org/10.1186/s13054-019-2563-x>.
47. Zimmerman LP, Reyfman PA, Smith ADR, Zeng Z, Kho A, Sanchez-Pinto LN, et al. Early prediction of acute kidney injury following ICU admission using a multivariate panel of physiological measurements. *BMC Med Inform Decis Mak.* 2019;19:16. <https://doi.org/10.1186/s12911-019-0733-z>.

48. Sun M, Baron J, Dighe A, Szolovits P, Wunderink RG, Isakova T, et al. Early prediction of acute kidney injury in critical care setting using clinical notes and structured multivariate physiological measurements. *Stud Health Technol Inform.* 2019;264:368–72. <https://doi.org/10.3233/shti190245>.
49. Gong K, Lee HK, Yu K, Xie X, Li J. A prediction and interpretation framework of acute kidney injury in critical care. *J Biomed Inform.* 2021;113:103653. <https://doi.org/10.1016/j.jbi.2020.103653>.
50. Xu Z, Luo Y, Adekkannattu P, Ancker JS, Jiang G, Kiefer RC, et al. Stratified mortality prediction of patients with acute kidney injury in critical care. *Stud Health Technol Inform.* 2019;264:462–6. <https://doi.org/10.3233/shti190264>.
51. Morid MA, Sheng ORL, Fiol GD, Facelli JC, Bray BE, Abdelrahman S. Temporal pattern detection to predict adverse events in critical care: case study with acute kidney injury. *JMIR Med Inform.* 2020;8:e14272. <https://doi.org/10.2196/14272>.
52. Parreco J, Soe-Lin H, Parks JJ, Byerly S, Chattoor M, Buick JL, et al. Comparing machine learning algorithms for predicting acute kidney injury. *Am Surg.* 2019;85:725–9. <https://doi.org/10.1177/000313481908500731>.
53. Chiofalo C, Chbat N, Ghosh E, Eshelman L, Kashani K. Automated continuous acute kidney injury prediction and surveillance: a random forest model. *Mayo Clin Proc.* 2019;94:783–92. <https://doi.org/10.1016/j.mayocp.2019.02.009>.
54. Wang Y, Wei Y, Yang H, Li J, Zhou Y, Wu Q. Utilizing imbalanced electronic health records to predict acute kidney injury by ensemble learning and time series model. *BMC Med Inform Decis Mak.* 2020;20:238. <https://doi.org/10.1186/s12911-020-01245-4>.
55. Ibrahim NE, McCarthy CP, Shrestha S, Gaggin HK, Mukai R, Magaret CA, et al. A clinical, proteomics, and artificial intelligence-driven model to predict acute kidney injury in patients undergoing coronary angiography. *Clin Cardiol.* 2019;42:292–8. <https://doi.org/10.1002/clc.23143>.
56. Huang C, Murugiah K, Mahajan S, Li S-X, Dhruba SS, Haimovich JS, et al. Enhancing the prediction of acute kidney injury risk after percutaneous coronary intervention using machine learning techniques: a retrospective cohort study. *PLoS Med.* 2018;15:e1002703. <https://doi.org/10.1371/journal.pmed.1002703>.
57. Huang C, Li S-X, Mahajan S, Testani JM, Wilson FP, Mena CI, et al. Development and validation of a model for predicting the risk of acute kidney injury associated with contrast volume levels during percutaneous coronary intervention. *JAMA Netw Open.* 2019;2:e1916021. <https://doi.org/10.1001/jamanetworkopen.2019.16021>.
58. Tseng P-Y, Chen Y-T, Wang C-H, Chiu K-M, Peng Y-S, Hsu S-P, et al. Prediction of the development of acute kidney injury following cardiac surgery by machine learning. *Crit Care.* 2020;24:478. <https://doi.org/10.1186/s13054-020-03179-9>.
59. Rank N, Pfahringer B, Kempfert J, Stamm C, Kühne T, Schoenrath F, et al. Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance. *NPJ Digit Med.* 2020;3:139. <https://doi.org/10.1038/s41746-020-00346-8>.
60. Thottakkara P, Ozrazgat-Baslanti T, Hupf BB, Rashidi P, Pardalos P, Momcilovic P, et al. Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. *PLoS One.* 2016;11:e0155705. <https://doi.org/10.1371/journal.pone.0155705>.
61. Adhikari L, Ozrazgat-Baslanti T, Ruppert M, Madushani RWMA, Paliwal S, Hashemighouchani H, et al. Improved predictive models for acute kidney injury with IDEA: intraoperative data embedded analytics. *PLoS One.* 2019;14:e0214904. <https://doi.org/10.1371/journal.pone.0214904>.
62. Lei VJ, Luong T, Shan E, Chen X, Neuman MD, Eneanya ND, et al. Risk stratification for postoperative acute kidney injury in major noncardiac surgery using preoperative and intraoperative data. *JAMA Netw Open.* 2019;2:e1916921. <https://doi.org/10.1001/jamanetworkopen.2019.16921>.
63. Jeon N, Staley B, Henriksen C, Lipori GP, Winterstein AG. Development and validation of an automated algorithm for identifying patients at higher risk for drug-induced acute kidney injury. *Am J Health Syst Pharm.* 2019;76:654–66. <https://doi.org/10.1093/ajhp/zxz043>.
64. Martinez DA, Levin SR, Klein EY, Parikh CR, Menez S, Taylor RA, et al. Early prediction of acute kidney injury in the emergency department with machine-learning methods applied to electronic health record data. *Ann Emerg Med.* 2020;76:501–14. <https://doi.org/10.1016/j.annemergmed.2020.05.026>.
65. Weisenthal SJ, Quill C, Farooq S, Kautz H, Zand MS. Predicting acute kidney injury at hospital re-entry using high-dimensional electronic health record data. *PLoS One.* 2018;13:e0204920. <https://doi.org/10.1371/journal.pone.0204920>.
66. Park N, Kang E, Park M, Lee H, Kang H-G, Yoon H-J, et al. Predicting acute kidney injury in cancer patients using heterogeneous and irregular data. *PLoS One.* 2018;13:e0199839. <https://doi.org/10.1371/journal.pone.0199839>.
67. Sandokji I, Yamamoto Y, Biswas A, Arora T, Ugwuwo U, Simonov M, et al. A time-updated, parsimonious model to predict AKI in hospitalized children. *J Am Soc Nephrol.* 2020;31:1348–57. <https://doi.org/10.1681/asn.2019070745>.
68. Sutherland SM, Chawla LS, Kane-Gill SL, Hsu RK, Kramer AA, Goldstein SL, et al. Utilizing electronic health records to predict acute kidney injury risk and outcomes: workgroup statements from the 15th ADQI consensus conference. *Can J Kidney Health Dis.* 2016;3:99. <https://doi.org/10.1186/s40697-016-0099-4>.
69. Churpek MM, Carey KA, Edelson DP, Singh T, Astor BC, Gilbert ER, et al. Internal and external validation of a machine learning risk score for acute kidney injury.

- JAMA Netw Open. 2020;3:e2012892. <https://doi.org/10.1001/jamanetworkopen.2020.12892>.
70. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035. <https://doi.org/10.1038/sdata.2016.35>.
71. Lundberg S, Lee SI. A unified approach to interpreting model predictions. 2017. arXiv:1705.07874v2.
72. Ugwuowo U, Yamamoto Y, Arora T, Saran I, Partridge C, Biswas A, et al. Real-time prediction of acute kidney injury in hospitalized adults: implementation and proof of concept. Am J Kidney Dis. 2020;76:806–14. <https://doi.org/10.1053/j.ajkd.2020.05.003>.
73. Driest SLV, Wang L, McLemore MF, Bridges BC, Fleming GM, McGregor TL, et al. Acute kidney injury risk-based screening in pediatric inpatients: a pragmatic randomized trial. Pediatr Res. 2020;87:118–24. <https://doi.org/10.1038/s41390-019-0550-1>.



Oscar J. Pellicer-Valero, Carlo Barbieri, Flavio Mari, and
José D. Martín-Guerrero

Contents

Introduction	580
Anemia, Hemoglobin Prediction, and ESA Dosage Optimization	580
MPC-Based Approaches: Hemoglobin Prediction	581
Direct Dose Optimization	586
Comorbidities, Mortality Prediction, and Patient Clustering	588
Other Miscellaneous Applications	589
Future Developments and Conclusions	590
Cross-References	590
References	590

Abstract

The prevalence of end-stage renal disease is experiencing a relentless growth in modern societies, significantly reducing the quality of life of

the patients enduring it and imposing an increasingly unsustainable economic burden to global healthcare systems. Most patients with this condition undergo hemodialysis, with comorbidities such as anemia, cardiovascular risks, and mineral bone disorders being extremely common, hence imposing a further challenge for all involved actors. Artificial Intelligence has already shown an enormous potential at solving key problems in these areas. In particular, anemia management has received the greatest attention, due to the extremely high cost of the erythropoietin-stimulating agents that are employed to treat it, the serious complications of their misadministration, and the huge inter- and inter-patient variability of the treatment. As such, several AI-based systems for dose optimization have been proposed and even commercialized over the years. A thorough presentation of the

Supplementary Information: The online version of this chapter (https://doi.org/10.1007/978-3-030-64573-1_254) contains supplementary material, which is available to authorized users.

O. J. Pellicer-Valero (✉) · J. D. Martín-Guerrero
Intelligent Data Analysis Laboratory, Department of
Electronic Engineering, ETSE (Engineering School),
Universitat de València (UV), Bujassot, Valencia, Spain
e-mail: Oscar.Pellicer@uv.es; jose.d.martin@uv.es

C. Barbieri · F. Mari
Fresenius Medical Care, Bad Homburg, Germany
e-mail: Carlo.Barbieri@fmc-ag.com;
flavio.mari@fmc-ag.com

developments in this area will be tackled in this chapter, including a description of the Artificial Intelligence models empowering them, such as artificial neural networks – both feed-forward and recurrent ones – or Reinforcement Learning-based approaches. Additionally, further interesting applications of Artificial Intelligence to the field of hemodialysis will be discussed, such as patient clustering and comorbidity and mortality prediction, among others. AI is bound to revolutionize the field of hemodialysis, and the aim of this chapter is to present a snapshot of the most current developments, finishing with a projection of these trends into the future.

Keywords

End-stage renal disease · Hemodialysis · Erythropoietin-stimulating agents · Anemia control · Hemoglobin prediction · Artificial Intelligence · Reinforcement Learning · Deep Learning · Recurrent neural network · Clustering

Introduction

The prevalence of renal diseases has experienced a continued growth in the last decades mainly due to generalized population aging and the global epidemic of the metabolic syndrome (abdominal obesity, insulin resistance, hypertension, and hyperlipidemia). Affecting 14.5% of patients aged 65+ years [1], chronic kidney disease (CKD) slowly deteriorates the kidneys until end-stage renal disease (ESRD) is reached, a condition which requires very expensive renal replacement therapies (RRTs), such as hemodialysis (HD) or transplant. Among patients undergoing HD (86.9% of ESRD patients), most of them (98%) do in-center HD, where they usually have 3–5 hour-long sessions three times a week.

Even if ESRD prevalence is just 0.2% of the total US population, it represents more than 10% of the total Medicare spending. In Europe, both ESRD prevalence and expenditures are slightly lower, yet the problem remains the same [2]. Nowadays, healthcare systems around the world are faced with the difficult challenge of optimizing

the expenditure in RRTs and ensuring its sustainability for an ever-increasing number of patients, and no easy solution is in sight.

Furthermore, HD replaces the blood filtration function of the kidney to some extent, but the renal endocrine role cannot be replaced, and hence multiple comorbidities appear, such as anemia, cardiovascular diseases, mineral bone disorders, etc. for which the patient has to take an average of 10–12 different tablets per day.

The rest of this chapter is organized as follows: in section “[Anemia, Hemoglobin Prediction, and ESA Dosage Optimization](#),” the problem of anemia secondary to HD is discussed. This is the most fruitful research topic regarding AI in HD, with several AI-based commercial systems for anemia management being currently employed in clinical practice. Section “[Comorbidities, Mortality Prediction, and Patient Clustering](#)” will discuss AI applications for analyzing comorbidities, predicting mortality, and clustering patients into groups in order to gain insights about their most usual profiles. Finally, section “[Other Miscellaneous Applications](#)” will introduce some miscellaneous topics of research in the field. All AI algorithms relevant to the field will be introduced progressively as they appear in the literature, so a sequential read of the chapter is recommended. The chapter ends in section “[Future Developments and Conclusions](#)” with some concluding remarks and the authors’ view about the future of the field.

Anemia, Hemoglobin Prediction, and ESA Dosage Optimization

Anemia, which is defined as a decrease in red blood cell (RBC) or hemoglobin (Hb) concentration, is one of the most common comorbidities in CKD, with a prevalence of around 50% in CKD patients at any stage, and around 75% in ESRD patients [3]. The lowered number of RBCs and/or Hb (which is the protein inside RBCs which carries the oxygen) leads to a reduction in the oxygen transport, which causes tiredness and weakness, and has a profound impact in the quality of life (QoL) of the patient. Furthermore, it increases morbidity and mortality from cardiovascular diseases (which, in turn, further deteriorate the renal function).

Even if CKD secondary anemia (CKD-anemia) arises from a variety of causes (iron deficiency, gastrointestinal bleeding, shortened RBC survival, etc.), the most significant and specific cause is the reduction in erythropoietin (EPO) synthesis. EPO is a glycoprotein secreted by the kidney that is crucial for the erythropoiesis process, by which hematopoietic stem cells produced in the bone marrow differentiate into mature RBCs ending in the bloodstream.

In the 1990s, the appearance of exogenous erythropoiesis-stimulating agents (ESAs) revolutionized the anemia treatment, improving the management of CKD-anemia along with the QoL of the patient, and replacing RBC transfusions almost completely. However, the response to ESAs has very strong inter- and intra-patient variability, combined with nonlinear, time-dependent dynamics, which are complex to model and not yet fully understood [4]. Furthermore, current guidelines agree that ESA treatment should only aim at a partial correction of Hb levels, as there is plenty of clinical evidence of higher mortality rates associated with higher ESA dosage [5]. As such, a target of 10–12 g/dL is currently recommended [6], even if normal Hb levels lie above 13 g/dL in men, and 12 g/dL in women for the average population. Finally, although multiple exogenous ESAs have been developed along the years (namely, epoetin alfa, darbepoetin beta, MPG-epoetin beta, etc.), their cost is still high, and their economic burden is significant.

Ideally, ESA administration protocols should be patient-specific and aim at achieving the Hb targets while keeping dosage as low as possible, thereby reducing both healthcare costs and patient risks. However, such task is extremely difficult to perform manually in clinical practice, and no universally accepted guidelines exist. This has led to the development of a huge variety of models to automatize the task, which can be divided in two main categories: model predictive control (MPC)-based approaches and direct ESA dose optimization.

MPC-Based Approaches: Hemoglobin Prediction

MPC-based approaches are the most common in the literature, and all currently available

commercial systems are based upon these. The core component of an MPC is the model, which should be able to accurately predict future Hb values based on an ESA dose proposal, patient characteristics, and current Hb values, among others. Then, the optimizer component takes a Hb reference to be achieved as input and uses the model to suggest the optimal treatment plan as an output. The Hb prediction model can be seen as a surrogate for the patient, with which a naïve optimizer could simply try several competing treatment plans in silico and choose the best one. Since the speed of the control is not an issue, such naïve controllers are commonly employed for this application, and the research focus is instead shifted toward finding the best possible Hb models. In practice, commercial systems for anemia control work as decision support systems, which suggest an ESA dose that the physician is then able to either follow or ignore. Figure 1 shows a schema of an MPC controller for Hb prediction.

There are two main groups of models for Hb prediction: physiologically based models and Artificial Intelligence (AI)-based models.

Physiologically Based Models

These models usually consist in a set of differential equations that describe the dynamics of erythropoiesis from first principles. The earliest of such models can be attributed to [7]. However, most current works are based upon [8], with subsequent modifications including further interactions within the erythropoiesis process, such as non-linear cell maturation and hormone decay rate [9] or neocytolysis [10].

The parameters of physiologically based models (e.g., RBC lifespan) are typically adjusted using the first months of a patient's history, *the descriptive phase*, after which the model can be employed for actual Hb prediction, *the prescriptive phase* [11]. The adjustment process is usually performed using direct search methods, e.g., trying random sets of parameters, and keeping the set for which the simulation best resembles the patient history during the descriptive phase. The physiologically based Mayo Clinic Anemia Management System (MCAMS) developed by [11, 12] and licensed to Physician Software

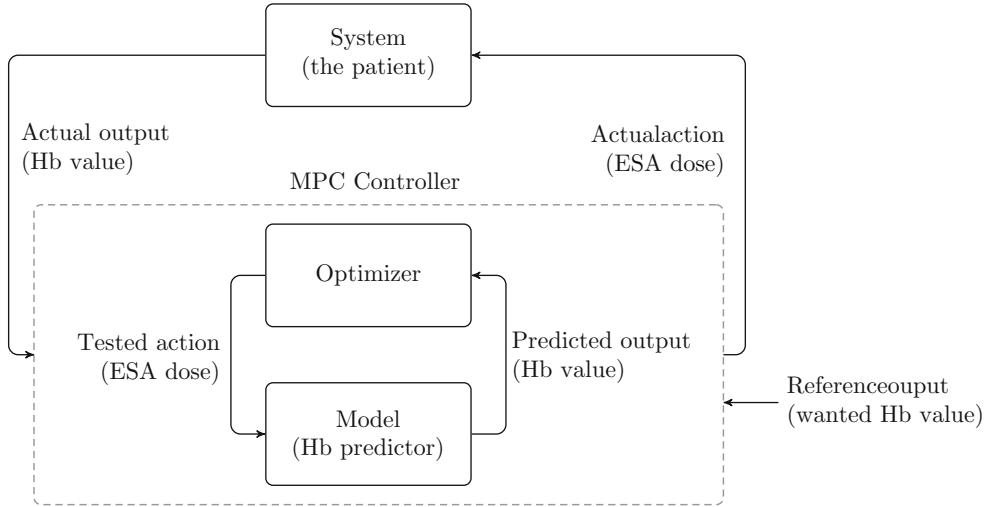


Fig. 1 MPC controller for Hb prediction: a naïve optimizer could use the model to test several possible actions (ESA doses), and simply choose the one which achieves a predicted Hb that is closest to the reference (the desired Hb value)

Systems has shown to achieve improvements in the percentage of patients within target Hb values.

AI-Based Models

AI-based models do not stem from an underlying erythropoiesis model; instead, they try to learn one purely from retrospective data in which the values of both the independent variable and the dependent variable (corresponding with the outcome of the model) are known. In AI research, such models are called supervised learning models, and they are based on the diagram presented in Fig. 2.

As it can be seen, the model itself is just a mathematical function $f_\theta(x)$, which takes an input x (e.g., ESA dose, patient characteristics, current Hb value, etc.) and produces an output \hat{y} (e.g., predicted Hb value in 30 days). The function $f_\theta(x)$ is parametric, which means that it can represent many different input-output relationships depending on its parameters θ , which are just numbers. During the learning phase, the model is shown many $\langle x, y \rangle$ pairs (where y is the ground truth for the output), and the model updates the value of its parameters in such a way that the predicted output \hat{y} and the ground truth y are as similar as possible for any given input x .

Linear Regression

Linear regression (LR) is one of the simplest supervised learning models. In [13], the authors

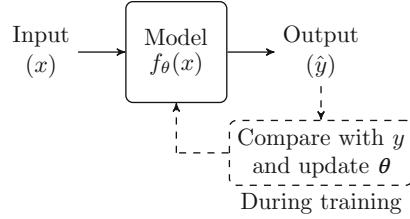


Fig. 2 General supervised learning model

use fuzzy sets to combine the output (predicted Hb) from three LR models, where each model corresponds to a prototypical patient according to their ESA sensitivity and average transferrin saturation (TSat) levels. The LR model used by [13] has been reproduced in Eq. 1:

$$\begin{aligned}
 Hb_{t+1} &= f_\theta(x) \\
 &= \theta_0 + \theta_1 \cdot EPO_{t-1} + \theta_2 \cdot EPO_t \\
 &\quad + \theta_3 \cdot EPO_{t+1} + \theta_4 \cdot Hb_{t-1} \\
 &\quad + \theta_5 \cdot Hb_t + \theta_6 \cdot TSat_t
 \end{aligned} \tag{1}$$

where t is the current time and $t-1, t, t+1$ denote that a value corresponds to the previous, current, or next month, respectively. Connecting with the general diagram in Fig. 2, the output $\hat{y} = Hb_{t+1}$, the input $x = [EPO_{t-1}, EPO_t, EPO_{t+1}, Hb_{t-1}, Hb_t, TSat_t]$, the parameters $\theta = [\theta_0, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6]$, and the model $f_\theta(x)$ correspond with the right-hand side of Eq. 1. Hence, the authors proposed to model the Hb

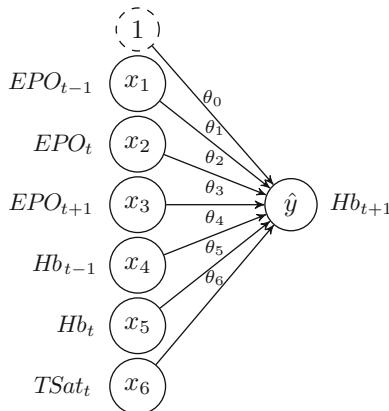


Fig. 3 Graphical representation of the linear regression model from [13]

in the next month (Hb_{t+1}) as a linear combination of previous (EPO_{t-1}), current (EPO_t), and future/proposed EPO doses (EPO_{t+1}), previous (Hb_{t-1}) and current Hb values (Hb_t), and current TSat ($TSat_t$). The parameters θ were learned automatically from medical records of 186 patients; this is a very powerful feature of AI algorithms, since no previous knowledge must be known about the problem (unlike with physiologically based models), they rather discover it by looking at the data. A graphical representation of the LR model from Eq. 1 is shown in Fig. 3.

It must be noted that, formally, the model in Eq. 1 is known an autoregressive model with exogenous inputs (ARX), since Hb_{t+1} depends on previous values of itself (autoregressive), and it includes other (exogenous) inputs.

Feed-Forward Neural Networks

The main disadvantage of LR is its very limited modeling power. Feed-forward neural networks (FFNNs) (also called multilayer perceptrons (MLPs)) are the most common kind of artificial neural network (ANN). Unlike LR, FFNNs are able to model very complex input-output relationships, including interactions among the inputs and arbitrary nonlinearities. For the task of Hb prediction, ANNs were used for the first time in [14, 15], where the authors demonstrated that they could yield more accurate results than ARX and radial basis function (RBF) networks. Since then, many more have appeared over the years, such as [16], where a significant jump in the number of patients (more than 10,000) and the number of inputs

(more than 20) was introduced; [17], where the prediction horizon was extended to 3 months into the future, instead of just one; or [18], which describes the model that is currently implemented in the commercial Anemia Control Module (ACM) developed by Fresenius Medical Care, which has shown to improve anemia target achievement while reducing EPO-dose administration [19]. A similar model [20] with similarly positive results was licensed by Dosis Inc. as Strategic Anemia Advisor (SAA).

FFNNs can be thought of as a stack of standard LRs with nonlinear activation functions in-between. To better illustrate this idea, Fig. 4 shows a representation of a FFNN very similar to the one used in [18]. A FFNN has an input layer and an output layer (exactly as in LR), but it adds hidden layers between them, each layer being comprised of several neurons (also called units). For instance, the FFNN in Fig. 4 has two hidden layers with six neurons each (plus a bias neuron, represented by a 1). The study of ANNs with many hidden layers makes up the field of Deep Learning (DL).

Each neuron in a FFNN behaves exactly as a LR plus a nonlinearity. For instance, the value at neuron z_1^1 can be calculated according to Eq. 2, where a LR over inputs $x_1, x_2, \dots, x_7, x_8$ has been performed and a *ReLU* nonlinear activation function ($ReLU(a) = \max(0, a)$) has been applied to the result, as denoted by the mirrored L symbol in some of the neurons of Fig. 4. Adding an activation function (such as *ReLU*, *sigmoid*, or *tanh*) to each neuron is crucial; otherwise, all the operations would remain linear, and the whole FFNN would be equivalent to a single LR over the inputs. Equations very similar to Eq. 2 can be defined for all neurons in the FFNN, hence defining the mathematical behavior of the whole network:

$$\begin{aligned} z_1^1 &= ReLU \\ (\theta_{01}^1 \cdot 1 + \theta_{11}^1 \cdot x_1 + \theta_{21}^1 \cdot x_2 + \dots + \theta_{71}^1 \cdot x_7 + \theta_{81}^1 \cdot x_8) \end{aligned} \quad (2)$$

In all kinds of ANNs (including LR, since it can be seen as the simplest kind of ANN), the parameters θ are usually obtained by a procedure

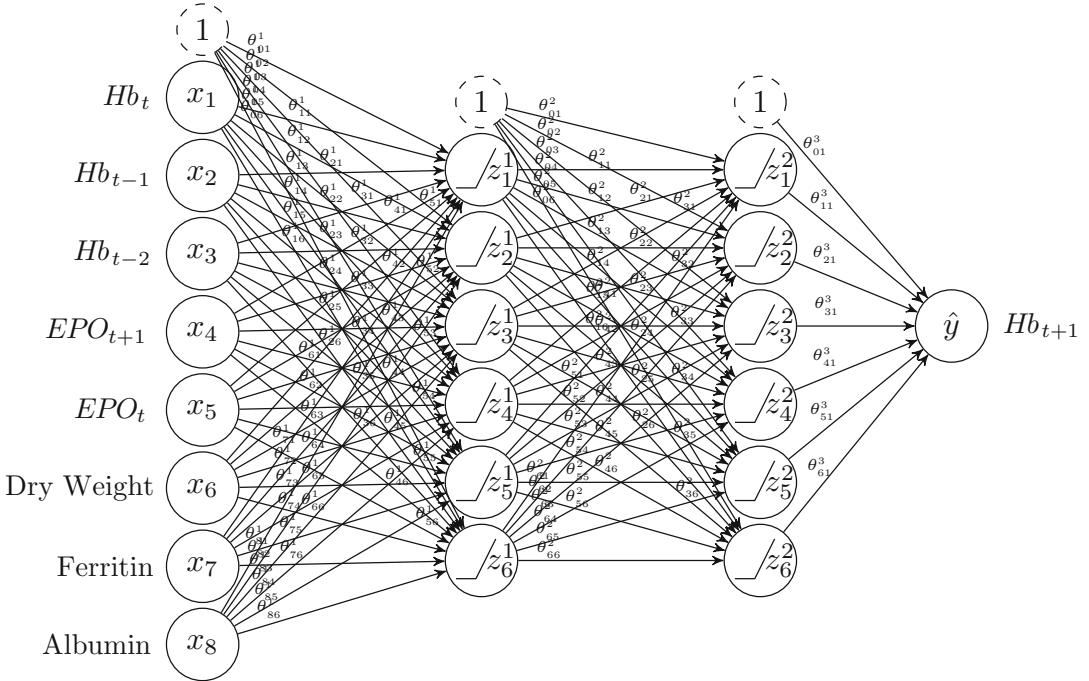


Fig. 4 FFNN for Hb prediction similar to the one in [18]

known as stochastic gradient descent (SGD). It starts by assigning random values to all parameters θ . First, a new sample $\langle x, y \rangle$ is taken, and the output \hat{y} of the ANN is calculated according to equations similar to Eq. 2. Second, the squared error between the ground truth y and the prediction \hat{y} is calculated as $(y - \hat{y})^2$. Third, the value of the parameters is updated in order to minimize that error. This process is repeated, iterating over the whole dataset multiple times, until the average error over all samples cannot be further improved.

Coming back to the Hb prediction model described in [18], the authors trained a FFNN on a dataset of 4,135 patients, with a total of 101,918 records from Italy, Spain, and Portugal. Moreover, they used a wide variety of inputs (also known as features), such as height, dry body weight, Hb, OcmKtV, ferritin, albumin, leucocytes, ESA dose (intravenous (IV) darbepoetin alpha), and iron dose (IV iron sucrose or IV iron gluconate), with some of them sampled at several points in the past (e.g., Hb_t , Hb_{t-1} , Hb_{t-2}). On one hand, using a high number of records from different countries exposes the ANN to a much more varied set of

data, making the trained ANN more general, and better able to deal with atypical, unseen situations. On the other hand, using many input features provides the ANN with more information about the patient, with which it can automatically learn better patient-dependent responses to ESA. Comparing with [13], where three LR models describing three prototypical patients were manually defined, FFNNs are able to learn the prototypes (in a sense) by themselves only by looking at the data.

ANNs are extremely powerful algorithms, able to learn very complex input-output relationships. In fact, a big FFNN could theoretically learn any dataset by heart, but then it would provide very poor predictions on data that it has not seen during training, which is a problem known as overfitting. To assess this issue, data is typically split into three sets: the training set contains around 70% of the data, and is used to train the ANN by updating its parameters with SGD; the validation set contains around 15% of the data, and is employed to assess the impact of changes to hyperparameters (such as number of hidden layers, number of neurons per

layer, etc.); finally, the test set is kept secret until the end, when the actual prediction performance of the ANN is eventually tested. For instance, in [18] the authors split the patients in proportions of 66%, 17%, and 17% for the training, validation, and test set, respectively, achieving similar performances in terms of mean absolute error (MAE) for all three sets, and hence proving that no overfitting occurred, a crucial issue to ensure a successful implementation of the model in daily clinical practice.

Recurrent Neural Networks

Despite the success of FFNNs in the task of Hb prediction, they are arguably not the best tool for the problem, since they have no notion of time despite the inputs (the patient history) being a time series. Fortunately, there is a kind of ANN known as recurrent neural network (RNN) that is specifically designed for time series modeling. The first instance where RNNs were applied to the problem of Hb prediction dates back to 2003 [14]; however, not much improvement was observed when compared with a simpler FFNN. It was not until 2020 when a sudden renewed interest in the topic appeared [21–23].

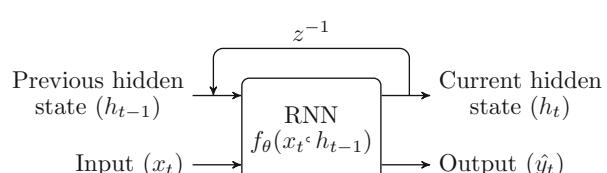
A high-level representation of a RNN can be found in Fig. 5. RNNs deal with sequential data (e.g., a medical history of a patient) by looking at one sample (one record: x_t) at a time. Then, they produce a prediction \hat{y}_t , and update their hidden state h_t , which is simply a vector of numbers that encodes information about all previous inputs $x_{t-1}, x_{t-2}, \dots, x_0$. When a new record from the same patient is input to the RNN, it can use the previous hidden state h_{t-1} along with the current information x_t to produce the Hb prediction \hat{y}_t . The hidden state can be understood as a memory of all the previous values and patterns of a patient's history that are useful for the RNN to predict the Hb.

Fig. 5 Overview of a recurrent neural network

In FFNNs, time is usually accounted for by adding the three last months of Hb values and ESA doses directly to the input (e.g., features Hb_t, Hb_{t-1}, Hb_{t-2} in Fig. 4). By contrast, RNNs encode time inherently by updating the hidden state. Hence, all previous information seen by the RNN could theoretically be remembered, including Hb values more than 3 months in the past (e.g., Hb_{t-3} , or even Hb_{t-20}), past values of other features (e.g., $Ferritin_{t-2}$), or, in general, any past information that could be useful for predicting Hb. Furthermore, since a prediction \hat{y} is provided at every time step t , RNNs can be used to predict Hb values from the very first month, without having to wait many months to adjust the parameters as with physiologically based models, or several months (usually three) to have all the required input data as with FFNN-based models.

RNNs can be comprised of different kinds of recurrent units (or neurons). The simplest recurrent unit is the Elman unit, which was used by [14]. However, Elman units are very limited, and are unable to keep long-term memories. These issues are solved by modern recurrent units, such as the long short-term memory (LSTM) units, employed by [21], and the gated recurrent units (GRUs), employed by [22, 23].

Perhaps, the most ambitious RNN model for Hb prediction to date is the one proposed by [23]. It was trained on more than 110,000 patients (around 3,000,000 total records) from clinics of 12 different countries, employing more than 30 different input features (including 3 kinds of ESA with both intravenous and subcutaneous administration). Despite (or maybe thanks to) the huge variability in the input data, it achieved a perceptible performance improvement when compared to previous state of the art while allowing to bring the benefits of algorithmic anemia control to virtually every patient without restriction (e.g.,



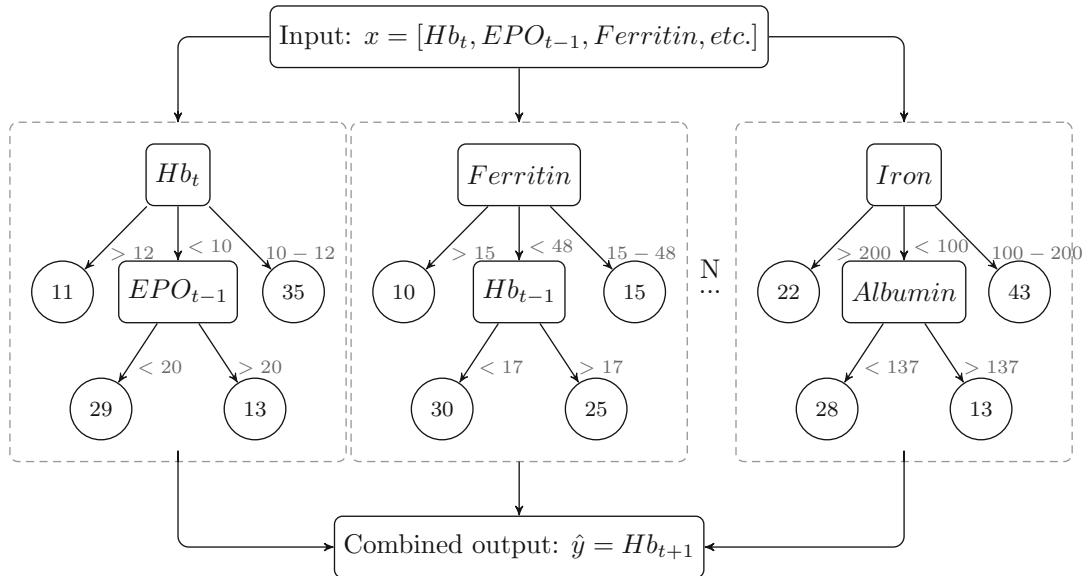


Fig. 6 Example of random forest for the problem of hemoglobin prediction

3 months of Hb history is no longer needed to obtain a prediction).

Other Supervised Learning Models

Despite FFNNs (and, more recently, RNNs) being the most popular for Hb prediction, many other models have also been explored over the years. In [15] the authors explore the use of radial basis function (RBF) networks, which are like FFNNs, but with parametric activation functions that compute a measure of the similarity between the inputs and a given prototypical input (patient). In [24] a modified version of support vector regression (SVR) with linear kernel is employed; this is similar to LR, but instead, SVR does not learn from points for which the prediction error is below a threshold. Finally, in [25], a model based on extremely randomized trees (ERTs) is used.

ERTs are a kind of random forest (RF), which is an ensemble method using decision trees (DTs). Ensemble methods combine the predictions from several individual models (e.g., by averaging), in order to improve the global performance. In a RF, DTs are trained independently on the same problem, but using a potentially different set of input features, tree architecture, or a different subset of the training samples. A toy example of a RF is

shown in Fig. 6, where the output of several DTs is combined to form the final output. As can be seen, when given an input, a trained DT takes a branch depending on the value of the input features, eventually arriving at a single final predicted value (represented as a circle).

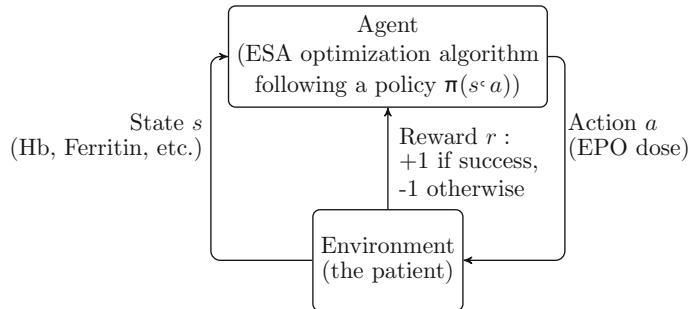
Direct Dose Optimization

Unlike MPC-based approaches, direct dose optimization systems are able to directly suggest the optimal ESA dose, without necessarily requiring an explicit model for Hb prediction. It must be noted that some of these methods do employ a Hb prediction model, but it is either extremely simple, or it is only used for the training of the dose optimizer, and not needed afterward.

Rule-Based Systems

Rule-based systems are derived from paper protocols designed by physicians that are expert in the topic. Authors in [26] proposed a rule-based system using the theory of fuzzy control, which was very popular in the 1990s, as it allowed to translate qualitative control rules into fuzzy sets that can be used to decide on the magnitude of a

Fig. 7 Typical Reinforcement Learning setup adapted to the problem of ESA optimization



control action; however, the model was not validated in real patients. In [27], the authors proposed a simple set of condition-action rules, which are guided mainly by the hematocrit (Hct) values (computed as three times the Hb value). For instance, one of such rules is: “If the calculated Hct is $\geq 39\%$, discontinue EPO and check the calculated Hct weekly. Resume EPO at 25% less than the previous dose as soon as the calculated Hct is $\leq 37.4\%$. ” The authors found no difference in Hb target achievement or ESA dosing when compared with manual control in real patients, although it is argued that the savings in terms of staff resources might be worth it. In [28], an analogous approach was followed, obtaining an improvement in Hb target achievement, which increased from 56% to 66% in 214 test patients.

Reinforcement Learning

Reinforcement Learning (RL) algorithms aim to learn a control policy $\pi(s, a)$ (an ESA dose optimization algorithm) that given a state s_t (e.g., $s_t = [EPO_t - 1, EPO_t, Hb_t - 1, Hb_t, TSat_t]$) chooses an action a_t (an ESA dose: $a_t = EPO_{t+1}$) that maximizes a long-term reward G_t (e.g., the longer that Hb levels stay within target range, the higher the value of G_t will be). G_t is usually defined as $G_t = r_t + \gamma \cdot r_{t+1} + \gamma^2 \cdot r_{t+2} + \dots$, where r_t is the reward that is obtained by choosing action a_t (e.g., $r_t = +1$ if Hb_{t+1} is in target range and $r_t = -1$ otherwise) and γ is a discount factor (e.g., $\gamma = 0.9$) that reduces the importance of future rewards (r_{t+1}, r_{t+2}, \dots). Finally, we can define the Q-function $Q_\pi(s, a)$ as the expected long-term reward G_t of choosing an action a at a given state s , and then following a policy π from there onward. An overview of the typical RL

setup adapted to the ESA optimization problem is shown in Fig. 7.

Instead of directly learning an optimal policy π^* , the objective of many RL algorithms is to approximate $Q_\pi(s, a)$ (with a table, a FFNN, an RBF, etc.) since, once it is known, the optimal policy π^* can always be found by taking the action a (from all possible actions in a state s) that maximizes $Q_\pi(s, a)$. This is

$$\pi^*(s) = \operatorname{argmax}_a Q(s, a) \quad (3)$$

There are two main groups of RL methods: on-policy (where the policy being learned is actually used to explore the environment) and off-policy (where the policy being learned does not coincide with the policy by which the environment is explored). In the context of ESA dose optimization, on-policy algorithms would require a patient (either real or simulated) to follow the policy that is proposed by the RL algorithm, which could be problematic in a real-life scenario; conversely, off-policy algorithms have the ability to learn an optimal policy from pre-acquired patient records (even if a sub-optimal ESA dosing policy was used). The most well-known on-policy RL algorithm is SARSA, which uses individual $\langle s_t, a_t, r_t, s_{t+1}, a_{t+1} \rangle$ tuples (hence its name) for learning the Q – function; and the most famous off – policy RL algorithm is Q – learning, which instead employs $\langle s_t, a_t, r_t, s_{t+1} \rangle$ tuples, where a_t might have been chosen following any policy. After learning an approximation to the Q-function, the optimal policy (the ESA dose optimizer) is trivial to obtain according to Eq. 3.

There exist a variety of papers applying these RL principles to ESA optimization. In [13], an off-policy algorithm known as $Q(\lambda)$ (very similar to the Q-learning algorithm) was trained on “virtual” patients, which were simulated by following the set of linear equations introduced as an example in section “[Linear Regression](#).[“](#) An RBF network was used to approximate the Q-function. Similarly, in a follow-up paper [29], the authors applied the on-policy RL algorithm SARSA over a set of patients that were simulated by a FFNN.

On-policy algorithms require an environment that can be freely explored (i.e., they require a patient on which arbitrary EPO doses can be probed). However, since this would be dangerous, in the previous papers, the authors decided to instead develop a simulated patient by using a supervised learning model for Hb prediction. Nevertheless, by training on simulated data, the RL algorithm depends entirely on how good the original Hb prediction model is, and, hence, its applicability in a real-life scenario, with real patient data, may be jeopardized. Conversely, in [30], an off-policy Q-learning algorithm was trained on data from real patients. The Q-function was approximated by both a table and a FFNN, achieving slightly better results with the latter approach. Finally, in [25], the off-policy fitted Q iteration algorithm was trained on augmented patient data (the number of patients was artificially increased by interpolating among the real patients), using an RBF network for approximating the Q-function. To assess the performance of the model, ERTs were used to fit a Hb prediction model, and the behavior of the proposed controller was simulated, revealing an improvement over standard EPO dosage protocol.

Overall, RL methods are a very promising technology, since they allow the development of optimal policies that can be learned directly from data, without the need of an explicit Hb prediction model. In other words, the RL philosophy is the following: let us focus on learning a good ESA optimizer, which might be a much simpler task than learning the underlying patient dynamics that govern erythropoiesis. However, unlike with MPC-based approaches, the decisions taken by RL methods cannot be easily understood and might be potentially dangerous if tested directly on real patients.

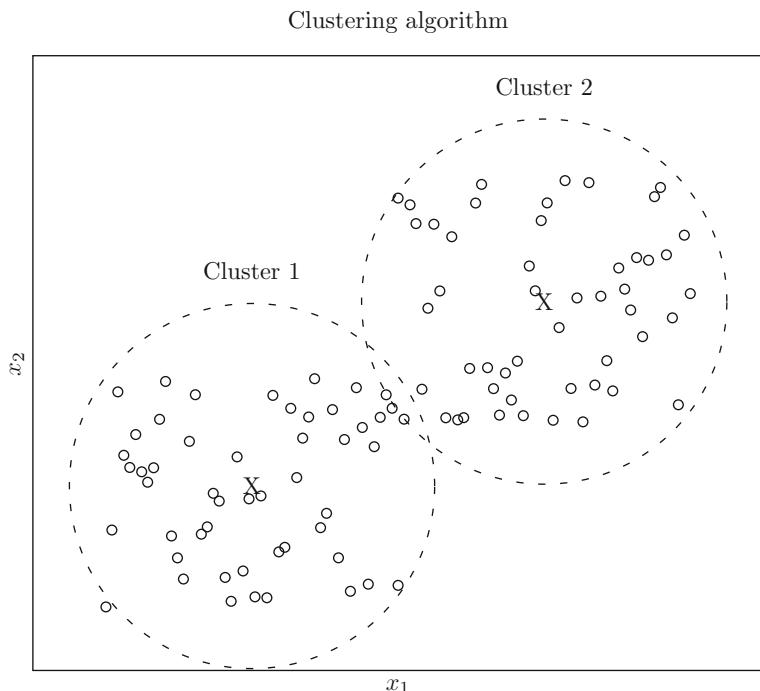
Comorbidities, Mortality Prediction, and Patient Clustering

Besides anemia control, the problem of early prediction of comorbidities associated with ESRD and HD, as well as patient (1-year, typically) mortality and patient clustering, has recently received some attention by the scientific community. Clustering algorithms are the most common form of unsupervised learning algorithms. Unlike supervised learning algorithms, these kinds of algorithms do not try to model an input-output relationship (see Fig. 2), but rather, they try to find relationships among the inputs themselves. A clustering algorithm, in particular, tries to find automatically groups of points (e.g., groups of patients), which are similar within their group, and different from others. Figure 8 shows a toy example of a clustering algorithm applied to two-dimensional data, where two clusters have been found.

Regarding patient clustering and mortality prediction, three articles have appeared in 2020. In [31], the authors trained a RF in 1,571 incident HD patients, achieving an area under the curve (AUC) of 0.728 for 1-year mortality prediction. AUC is a classification metric ranging from 0 to 1, with 1 meaning both perfect classification accuracy and absolute confidence by the model in the prediction. In [32], 101 incident HD patients were clustered by means of the hierarchical clustering algorithm, yielding 3 groups of patients (or clusters). The cluster where patients had (before HD) lower systolic blood pressures, lower serum creatinine, and urinary liver-type fatty acid-binding protein levels was associated with the highest 1-year mortality risk. Yet, the most ambitious study in this topic to date is [33], where almost 80,000 patients were clustered in 5 groups by the K-means algorithm. Then, a different support vector machine (SVM) classification model was trained for each of the clusters, achieving an AUC of 0.948 for 1-year mortality prediction. Furthermore, the characteristics of the patients inside each of the clusters revealed some interesting insights about the HD population and their prognosis.

In relation to comorbidities, a topic of great interest is the CKD-mineral bone disorder

Fig. 8 Example of clustering algorithm



(CKD-MBD), which is made patent by a drift from normal phosphate, calcium, or parathyroid hormone values, and is correlated with a significant mortality increase. Regulating all three parameters is extremely difficult due to their complex dynamics and interdependencies, and standard control protocols typically achieve variable degrees of success. In [34], an attempt was made at analyzing the associations among several of the variables that influence the problem by using RFs. This is still a very young field inside AI research, and publications on the topic are bound to appear in the coming years, due to the importance of the medical problem. Finally, in [35] a model for cardiovascular disease prediction in a 6-month window in incident HD patients was developed, achieving an AUC of 0.737 with a RF.

Other Miscellaneous Applications

There are several interesting AI applications in the HD field that do not belong to any specific category. For instance, [36] proposed a model for predicting arteriovenous fistula (AF) stenosis by processing AF sounds with a bidirectional LSTM convolutional neural network, which is a

supervised learning model designed specifically for processing signals (AF sounds) that are recorded over time (at several HD sessions). AF stenosis consists in the narrowing of the blood vessels used for the HD process. It is typically assessed by auscultation and assessed over time in a subjective manner. The system proposed by the authors aims at objectively assessing the stenosis (classifying it as either normal, hard, high, intermittent, or whistling), by processing the sound produced by the AF during HD. They achieved an AUC of 0.75–0.92 for the different classes, hence having the potential of becoming an objective measure of AF stenosis in clinical practice. Similarly, in [37], a FFNN-based model is used to predict session-specific Kt/V, fluid volume removal, heart rate, and blood pressure based on patient characteristics, historic hemodynamic responses, and dialysis-related prescriptions, achieving high success in predicting the first two values.

Regarding other topics, in [38], an assessment of performance trends in HD clinics was performed, by utilizing a kind of unsupervised learning algorithm known as self-organizing maps (SOMs) and Markov process theory. SOMs were also employed in [39] to identify the

characteristics of groups of patients according to their satisfaction with the HD treatment, hence allowing to explore areas of potential improvement. Finally, in [40], an exploration of the progress in wearable HD devices research is presented. The authors argue that AI will have a profound impact in such devices, which will have to be able to continually monitor the patient and respond to changes in homeostasis in an intelligent manner.

Future Developments and Conclusions

The relationship between HD and AI has been gaining momentum over the years, and it seems like this trend will continue into the foreseeable future, yielding results that will transcend the lab and revolutionize the clinical practice. As a clear example, several automatic ESA dose recommendation systems are already being used in real-life applications, and several others are likely on the way. These automated systems have the potential to bring personalized anemia management to a very large cohort of population, improving the QoL of the patients and reducing the costs of the increasingly unsustainable healthcare systems, especially in areas where resources are scarce and highly educated physicians are limited.

A classical (unfounded) concern with such systems is that they will replace doctors, but it is quite the opposite: AI will empower them. Just like any other medical device, AI systems are just tools that concentrate on doing one specific thing extremely well (e.g., ESA dose optimization), while the doctor is left with more time to focus on the patient as whole. In fact, if trained with enough data, AI systems will eventually outperform any human-based information-related process, since, unlike humans, AI models have the ability to easily extract and remember patterns from hundreds of thousands of patients. In that regard, the most limiting factor for AI development will be high-volume and high-quality data availability. Even if this is not a concern for a high-incidence condition such as CKD-anemia, it will definitely be for many other uncommon

diseases, and a global collaboration will likely be needed in such cases.

AI-based systems have a very promising journey ahead in helping HD patients live the best life possible, and it is expected that many more HD problems that still remain marginal in the field will be tackled very soon, including models for optimizing treatment and prediction of other CKD-associated comorbidities (such as CKD-MBD or cardiovascular diseases), systems to optimize the effectiveness and the experience of the patient during the HD process, or even the development of wearable artificial kidneys, which are still in very early stages, but are an extremely interesting technology nonetheless. AI has an unprecedented potential to revolutionize many medical fields for the better, and, as such, it should be understood and promoted by both scholars and physicians, with a shared objective of improving the life of the patients.

Cross-References

- [AIM in Electronic Health Records \(EHRs\)](#)
- [Artificial Intelligence in Kidney Pathology](#)
- [Artificial Intelligence in Medicine in Anemia](#)
- [Artificial Intelligence in Predicting Kidney Function and Acute Kidney Injury](#)

References

1. United States Renal Data System. 2019 USRDS annual data report: epidemiology of kidney disease in the United States. Bethesda; 2019.
2. Stel VS, Brück K, Fraser S, et al. International differences in chronic kidney disease prevalence: a key public health and epidemiologic research issue. *Nephrol Dial Transplant*. 2017;32:ii129–35. <https://doi.org/10.1093/ndt/gfw420>.
3. Thomas R, Kanso A, Sedor JR. Chronic kidney disease and its complications. *Prim Care Clin Off Pract*. 2008;35:329–44. <https://doi.org/10.1016/j.pop.2008.01.008>.
4. Chait Y, Kalim S, Horowitz J, et al. The greatly misunderstood erythropoietin resistance index and the case for a new responsiveness measure. *Hemodial Int*. 2016;20:392–8. <https://doi.org/10.1111/hdi.12407>.
5. U.S. Food and Drug Administration. FDA drug safety communication: modified dosing recommendations to

- improve the safe use of Erythropoiesis-Stimulating Agents (ESAs) in chronic kidney disease. FDA; 2011. <https://www.fda.gov/Drugs/DrugSafety/ucm259639.htm>. Accessed 17 Jan 2019.
6. Locatelli F, Bárány P, Covic A, et al. Kidney disease: improving global outcomes guidelines on anaemia management in chronic kidney disease: a European renal best practice position statement. *Nephrol Dial Transplant*. 2013;28:1346–59.
 7. Uehlinger DE, Gotch FA, Sheiner LB. A pharmacodynamic model of erythropoietin therapy for uremic anemia. *Clin Pharmacol Ther*. 1992;51:76–89. <https://doi.org/10.1038/clpt.1992.10>.
 8. Bélair J, Mackey MC, Mahaffy JM. Age-structured and two-delay models for erythropoiesis. *Math Biosci*. 1995;128:317–46. [https://doi.org/10.1016/0025-5564\(94\)00078-E](https://doi.org/10.1016/0025-5564(94)00078-E).
 9. Ackleh AS, Deng K, Ito K, Thibodeaux J. A structured erythropoiesis model with nonlinear cell maturation velocity and hormone decay rate. *Math Biosci*. 2006;204:21–48. <https://doi.org/10.1016/j.mbs.2006.08.004>.
 10. Fuertinger DH, Kappel F, Zhang H, et al. Prediction of hemoglobin levels in individual hemodialysis patients by means of a mathematical model of erythropoiesis. *PLoS One*. 2018;13:e0195918. <https://doi.org/10.1371/journal.pone.0195918>.
 11. Rogers J, Gallaher EJ, Dingli D. Personalized ESA doses for anemia management in hemodialysis patients with end-stage renal disease. *Syst Dyn Rev*. 2018;34: 121–53. <https://doi.org/10.1002/sdr.1606>.
 12. McCarthy JT, Hocum CL, Albright RC, et al. Biomedical system dynamics to improve anemia control with darbepoetin alfa in long-term hemodialysis patients. *Mayo Clin Proc*. 2014;89:87–94. <https://doi.org/10.1016/j.mayocp.2013.10.022>.
 13. Gaweda AE, Muezzinoglu MK, Aronoff GR, et al. Individualization of pharmacological anemia management using reinforcement learning. *Neural Netw*. 2005;18:826–34. <https://doi.org/10.1016/j.neunet.2005.06.020>.
 14. Martín Guerrero JD, Soria Olivas E, Camps Valls G, et al. Use of neural networks for dosage individualisation of erythropoietin in patients with secondary anemia to chronic renal failure. *Comput Biol Med*. 2003;33:361–73. [https://doi.org/10.1016/S0010-4825\(02\)00065-3](https://doi.org/10.1016/S0010-4825(02)00065-3).
 15. Gaweda AE, Jacobs AA, Brier ME, Zurada JM. Pharmacodynamic population analysis in chronic renal failure using artificial neural networks – a comparative study. *Neural Netw*. 2003;16:841–5. [https://doi.org/10.1016/S0893-6080\(03\)00084-4](https://doi.org/10.1016/S0893-6080(03)00084-4).
 16. Martínez-Martínez JM, Escandell-Montero P, Barbieri C, et al. Prediction of the hemoglobin level in hemodialysis patients using machine learning techniques. *Comput Methods Prog Biomed*. 2014;117: 208–17. <https://doi.org/10.1016/j.cmpb.2014.07.001>.
 17. Barbieri C, Bolzoni E, Mari F, et al. Performance of a predictive model for long-term hemoglobin response to Darbepoetin and Iron Administration in a Large Cohort of hemodialysis patients. *PLoS One*. 2016;11:e0148938. <https://doi.org/10.1371/journal.pone.0148938>.
 18. Barbieri C, Mari F, Stopper A, et al. A new machine learning approach for predicting the response to anemia treatment in a large cohort of end stage renal disease patients undergoing dialysis. *Comput Biol Med*. 2015;61:56–61. <https://doi.org/10.1016/j.combiomed.2015.03.019>.
 19. Barbieri C, Molina M, Ponce P, et al. An international observational study suggests that artificial intelligence for clinical decision support optimizes anemia management in hemodialysis patients. *Kidney Int*. 2016;90: 422–9. <https://doi.org/10.1016/j.kint.2016.03.036>.
 20. Gaweda AE, Aronoff GR, Jacobs AA, et al. Individualized anemia management reduces hemoglobin variability in hemodialysis patients. *J Am Soc Nephrol*. 2014;25:159–66. <https://doi.org/10.1681/ASN.2013010089>.
 21. Lobo B, Abdel-Rahman E, Brown D, et al. A recurrent neural network approach to predicting hemoglobin trajectories in patients with end-stage renal disease. *Artif Intell Med*. 2020;104:101823. <https://doi.org/10.1016/j.artmed.2020.101823>.
 22. Yoo T-H, Yun H-R, Chang JH. Development of Hemoglobin Prediction and Erythrocyte Stimulating Agent Recommendation Algorithm (HPERA) using recurrent neural network in end-stage kidney disease patients. *Nephrol Dial Transplant*. 2020;35 <https://doi.org/10.1093/ndt/gfaa142.p1374>.
 23. Pellicer-Valero OJ, Cattinelli I, Neri L, et al. Enhanced prediction of hemoglobin concentration in a very large cohort of hemodialysis patients by means of deep recurrent neural networks. *Artif Intell Med*. 2020;107 <https://doi.org/10.1016/j.artmed.2020.101898>.
 24. Martin-Guerrero JD, Camps-Valls G, Soria-Olivas E, et al. Dosage individualization of erythropoietin using a profile-dependent support vector regression. *IEEE Trans Biomed Eng*. 2003;50:1136–42. <https://doi.org/10.1109/TBME.2003.816084>.
 25. Escandell-Montero P, Chermisi M, Martínez-Martínez JM, et al. Optimization of anemia treatment in hemodialysis patients via reinforcement learning. *Artif Intell Med*. 2014;62:47–60. <https://doi.org/10.1016/j.artmed.2014.07.004>.
 26. Bellazzi R, Siviero C, Bellazzi R. Mathematical modeling of erythropoietin therapy in uremic anemia. Does it improve cost-effectiveness? *Haematologica*. 1994;79: 154–64.
 27. Miskulin DC, Weiner DE, Tighiouart H, et al. Computerized decision support for EPO dosing in hemodialysis patients. *Am J Kidney Dis*. 2009;54:1081–8. <https://doi.org/10.1053/j.ajkd.2009.07.010>.
 28. Lines SW, Lindley EJ, Tattersall JE, Wright MJ. A predictive algorithm for the management of anaemia in haemodialysis patients based on ESA pharmacodynamics: better results for less work. *Nephrol Dial Transplant*. 2012;27:2425–9. <https://doi.org/10.1093/ndt/gfr706>.

29. Gaweda AE, Muezzinoglu MK, Jacobs AA, et al. Model predictive control with reinforcement learning for drug delivery in renal Anemia management. In: 2006 International conference of the IEEE engineering in medicine and biology society. IEEE; 2006. p. 5177–80.
30. Martín-Guerrero JD, Gomez F, Soria-Olivas E, et al. A reinforcement learning approach for individualizing erythropoietin dosages in hemodialysis patients. *Expert Syst Appl.* 2009;36:9737–42. <https://doi.org/10.1016/j.eswa.2009.02.041>.
31. Garcia-Montemayor V, Martin-Malo A, Barbieri C, et al. Predicting mortality in hemodialysis patients using machine learning analysis. *Clin Kidney J.* 2020; <https://doi.org/10.1093/ckj/sfaa126>.
32. Komaru Y, Yoshida T, Hamasaki Y, et al. Hierarchical clustering analysis for predicting 1-year mortality after starting hemodialysis. *Kidney Int Reports.* 2020;5: 1188–95. <https://doi.org/10.1016/j.ekir.2020.05.007>.
33. Kanda E, Epureanu BI, Adachi T, et al. Application of explainable ensemble artificial intelligence model to categorization of hemodialysis-patient and treatment using nationwide-real-world data in Japan. *PLoS One.* 2020;15:e0233491. <https://doi.org/10.1371/journal.pone.0233491>.
34. Rodriguez M, Salmeron MD, Martin-Malo A, et al. A new data analysis system to quantify associations between biochemical parameters of chronic kidney disease-mineral bone disease. *PLoS One.* 2016;11: e0146801. <https://doi.org/10.1371/journal.pone.0146801>.
35. Ion Titapiccolo J, Ferrario M, Cerutti S, et al. Artificial intelligence models to stratify cardiovascular risk in incident hemodialysis patients. *Expert Syst Appl.* 2013;40:4679–86. <https://doi.org/10.1016/j.eswa.2013.02.005>.
36. Ota K, Nishiura Y, Ishihara S, et al. Evaluation of hemodialysis arteriovenous bruit by deep learning. *Sensors.* 2020;20:4852. <https://doi.org/10.3390/s20174852>.
37. Barbieri C, Cattinelli I, Neri L, et al. Development of an artificial intelligence model to guide the Management of Blood Pressure, fluid volume, and Dialysis dose in end-stage kidney disease patients: proof of concept and first clinical assessment. *Kidney Dis.* 2019;5:28–33. <https://doi.org/10.1159/000493479>.
38. Cattinelli I, Bolzoni E, Chermisi M, et al. Computational intelligence for the balanced scorecard: studying performance trends of hemodialysis clinics. *Artif Intell Med.* 2013;1:1–11. <https://doi.org/10.1016/j.jinf.2020.04.010>.
39. Martín Guerrero JD, Marcelli D, Soria-Olivas E, et al. Self-Organising maps: a new way to screen the level of satisfaction of dialysis patients. *Expert Syst Appl.* 2012;39:8793–8. <https://doi.org/10.1016/j.eswa.2012.02.001>.
40. Hueso M, Navarro E, Sandoval D, Cruzado JM. Progress in the development and challenges for the use of artificial kidneys and wearable Dialysis devices. *Kidney Dis.* 2019;5:3–10. <https://doi.org/10.1159/000492932>.



Artificial Intelligence in Public Health

42

Thomas Lefèvre and Sabine Guez

Contents

Introduction	594
Defining Public Health and Social Medicine: Toward Greater Precision, or a Regression from the Collective to the Individual?	595
Public Health Versus Individual Health, Social Medicine Versus Personalized Medicine	595
Precision Versus Individualization. Precision Public Health	595
Public Health, a Practice Based on Data, on Information, or on Evidence?	596
Where Is Artificial Intelligence Used in Public Health?	596
There Is No Artificial Intelligence Without Data: Data Federation and “New” Types of Data	596
The Citizen and the Patient: Producer, Actor, and Manager of Their Own Health	597
AI and Health Surveillance Systems	597
Learning Healthcare Systems (LHSs)	597
Risk and Insurance	598
Organization and Governance	599
Future Challenges	600
Challenges Shared with Other Health Fields	600

T. Lefèvre (✉)

IRIS Institut de Recherche Interdisciplinaire sur les enjeux Sociaux, UMR8156 CNRS – U997 Inserm – EHESS – Université Sorbonne Paris Nord, Paris, France

Department of Forensic and Social Medicine, AP-HP, Jean Verdier Hospital, Bondy, France
e-mail: thomas.lefeuvre@univ-paris13.fr

S. Guez
IRIS Institut de Recherche Interdisciplinaire sur les enjeux Sociaux, UMR8156 CNRS – U997 Inserm – EHESS – Université Sorbonne Paris Nord, Paris, France

More Specific Challenges	601
Conclusion and Outlook	601
Cross-References	601
References	601

Abstract

The potential contribution of artificial intelligence to public health is made up, in a sense, of its many specific contributions to each of the medical disciplines, in that each individual contribution will help to improve the health of the population by making more efficient use of the resources of the healthcare system and improving individual health. There are, however, a few specific areas of public health and social medicine in which AI could bring about an evolution, or even a revolution. This may be the case for precision public health, which draws on data that is both more varied in nature, such as behavioral data from connected objects, and more precise in temporal and spatial resolution. Public health uses aggregate, often macroscopic, data to make health policy decisions. Such indicators are imperfect; they can be inconsistent or misleading and are often most useful in retrospect, rather than at the desired moment. Evidence-based policy would appear to be more legitimate and robust. AI could also change the way public health systems are organized at various levels. Learning healthcare systems, for example, are designed to adapt more or less autonomously to changing health needs. In any case, the challenges of effectively using AI will arise in public health as elsewhere, and they may even be exacerbated or intensified by the well-known and unresolved tension between individual preferences and collective preferences.

Keywords

Public health · Artificial intelligence · Social medicine · Risk factor · Learning healthcare systems · Health insurance · Evidence-based

policy · Precision public health · Population health

Introduction

Public health can be a delicate matter, as it is a crosscutting and complementary discipline, rather than merely supplementary to other medical disciplines. Public health and the related field of social medicine serve somewhat as foils to the other medical disciplines, which are characterized by a biological and reductionist paradigm focused on the individual [1]: care is provided to people who may present with similar characteristics, but who are all unique. This makes it tempting to see public health and social medicine as somewhat of a supplement, adding context or a collective viewpoint to the footnotes of textbooks for other medical specialties. This can be seen in the many scientific articles that begin with: “This is a major public health issue because...” The collective dimension of public health means that it shares certain concepts and tools with epidemiology. Thus, at a time when artificial intelligence in medicine is often synonymous with P4 medicine (personalized, predictive, preventative, and participatory), there may be little specific overlap between public health and artificial intelligence. With this in mind, in this chapter we will discuss what is meant by public health more specifically: namely, all that which concerns the health of populations, the healthcare system, or, in other words, the collective and organizational dimensions of health [2]. We will look at the cases of social medicine and precision public health. The arrival of artificial intelligence in the field of public health, before it has even had the chance to bring about any evolutions or revolutions of note, has already

reactivated well-known and unresolved tensions between individual preferences and social preferences.

Defining Public Health and Social Medicine: Toward Greater Precision, or a Regression from the Collective to the Individual?

Public Health Versus Individual Health, Social Medicine Versus Personalized Medicine

Here, we use public health to mean the field concerned with population health, acting as a complement to medicine, which focuses on individual health [2]. Public health is primarily interested in prevention and how healthcare is organized. Population health is tied to the health of the individuals within that population, but the public health approach is intrinsically collective. Good examples might include infectious diseases and vaccination [3]. Infectious diseases that can be transmitted from human to human are a concern for both the individual and the group: there may be a need to create new institutions and networks, to mobilize certain techniques (vaccines), and to adopt policies to attempt to control or eradicate an epidemic at the population level, without a clear direct individual interest for each person. Infected people can pass on a virus without experiencing any ill effects, or with only minor symptoms. Yet, they could spread the disease to others who might develop symptoms or suffer more serious consequences. Public health looks at the organization of the healthcare system and healthcare professionals, at how public health policy is developed, and at the resources made available to implement these policies. It also studies the people involved at these different levels, based on the fundamental principle of health democracy that all citizens should be able to contribute to and give their opinion on public health decisions [4].

In the context of public health, social medicine can be understood as the application of public health principles to individual medical practice

[5]. Health practitioners must, of course, consider the unique and individual situation of each patient they see, but the patient's requests, needs, and resources should be understood in the context of the specific collective dimension of public health. In particular, this means considering the patient's environment, in the broadest possible sense, as well as their interactions with this environment, be they professional, socioeconomic, health, cultural, etc. Social medicine takes social determinants into account in its approach [6].

Medical and clinical practice primarily focuses on the direct relationship between the patient and the healthcare professional. The dominant paradigm today is biomedical and individual. Until now, the use of artificial intelligence in medicine has reflected the underlying principles of this paradigm and can be summarized under what is called P4 medicine: personalized, predictive, preventative, and participatory [7]. While "participatory" medicine may seem to have a collective dimension, this is quickly reduced to encouraging individuals to adopt responsible, informed, and autonomous behavior with regard to their own health. Thus, while it was not their original intention, most current and planned projects to use AI in medicine aim to augment doctors' usual capabilities in this patient-professional context: diagnostic, prognostic, and therapeutic capabilities [8]. The overall goal seems to be a more personalized and individualized form of care, with custom diagnosis and targeted treatments based on the patient's pharmacogenetic profile [9].

Precision Versus Individualization. Precision Public Health

Precision would therefore appear to be a key benefit of the use of AI in health. But what exactly is this precision? And what forms can it take? As we have just seen, it is widely held to mean the greater personalization and individualization of clinical and surgical procedures. This personalization, however, remains largely virtual, or at least similar to the use of classic risk factor analysis, applied to an individual: it says something about the patient, insofar as they are a member of a group

of people, large or small, that share a certain number of characteristics. AI can have a double benefit in this context, helping to reveal risk factors that are unknown or difficult to determine using traditional methods [8], and more precisely defining risk strata, thereby increasing patient group granularity. The collective aspect must therefore be taken into account upstream, during the statistical identification of characteristics of interest. However, it is less useful, and can even cause problems, when applying results derived from homogeneous groups to individual situations [10]. In public health, there is another kind of precision that may be applied.

Since the 2010s, the concept of precision public health has been discussed in important scientific and medical journals [11]. Precision can be assessed across all areas of public health: the precision of estimates and forecasts (number of cases, of deaths), spatial precision (indicators at the country level vs the kilometric or individual level), the precision of prevention information and messaging (undifferentiated advertising in the public space vs individual information, adjusted based on personal characteristics), precision in service provision (basing treatment costs on individual behaviors), and precision in decision-making and policy measures. For example, in Africa, recent estimates have indicated an overall decrease in infant mortality. However, an analysis with a 5 km resolution reveals sharply contrasting spatial inequalities, with infant mortality falling in some areas but actually rising in others [12].

Public Health, a Practice Based on Data, on Information, or on Evidence?

It cannot be said enough: there is no AI without data, whether that data is used to train algorithms or to apply them to real-world situations. Public health stands at the intersection between health and policy, and public health decision-making can reveal a lot about the state's prerogatives. Individual medicine is based on what is called "evidence-based medicine" [13], so the public health equivalent is "evidence-based policy," or EBP [14]. This emphasis on evidence is part of a

wider methodological framework. As big data and AI have taken off, the lines between data, information, and evidence in EBP have become blurred. Data is simply whatever is recorded or captured by some measuring device. Information can be based on data, when it is qualified and considered in a particular context that gives it value. Finally, evidence is information that comes from a process ensuring its soundness, over a long enough timescale to be relevant and after testing for any scientific errors, for example [15]. It remains difficult today to determine the added value of AI in its use of data, whether to produce information or evidence. The most fervent data supporters have called for data and AI to be used to guide decision-making without any theoretical or explanatory framework, or even without any human intervention [16]. This would lead us into what is called data-driven policy [17]. At the other extreme, traditional EBP models are situated downstream, with a focus on (human) expertise. There is also a third path, one that blends AI and expertise, a kind of evolution of older "expert systems."

We will now review several subfields of public health where AI may be used.

Where Is Artificial Intelligence Used in Public Health?

There Is No Artificial Intelligence Without Data: Data Federation and "New" Types of Data

The widespread and large-scale development of AI algorithms is predicated on increased access to relevant data. This is required to develop the algorithms, and then to move them into production, and finally to adapt them to working environments, interfacing them with other information systems. AI would seem to be a useful tool for organizing and federating all of this data, as well as processing and analyzing it. The diverse array of data sources presents an opportunity for public health, especially because we are so often lacking sufficient data: (i) real-life, repeated data, (ii) environmental data, (iii) behavioral data, and,

more specifically, (iv) data on exposure (which can be included under environmental data). AI in public health could therefore benefit from data from geographic information systems and more broadly from geospatial/geolocation data [18], data on digital footprints and individual activity, data from social networks and online forums [19], and data from connected objects, with a focus on passive data collection. These connected objects can include medical devices (implanted devices), as well as nonmedical devices, in particular smartphones.

The Citizen and the Patient: Producer, Actor, and Manager of Their Own Health

One of the best data sources for AI in public health, therefore, will be people themselves, whether they are patients or not, and whether they are sick or well. People are of interest whether or not they are presenting with a pathology, since the goal is to find associations between a certain exposure and a given health event. This exposure may be biological, or it may be social, cultural, economic, behavioral, etc. Every person therefore becomes a potential data “producer,” either actively or passively, while also becoming a “producer” and promoter of their own health. There is a loop that begins with the individual generating data, which is added to a pool of data collected from other individuals. This data pool is then processed using AI, and the original data “producer” may then be contacted with information or instructions about their health. A classic example would be the use of this model to improve therapeutic education [20].

Of course, AI can operate on a level well above that of the individual.

AI and Health Surveillance Systems

In the past, AI has been used in public health to monitor various potential threats, such as epidemics and the improper use of pharmaceuticals or medical devices [21]. The classic example is

Google Flu Trends, although it is no longer active and had certain clear limitations. Google Flu Trends claimed that it could predict flu outbreaks faster and more reliably than the CDC (Centers for Disease Control) by analyzing search terms entered into the Google search engine [22]. Various studies of the advantages and disadvantages showed that Google Flu Trends was not very precise and pointed out weaknesses in the algorithm, or rather algorithms, since the algorithm was replaced several times, although we do not know exactly how it was changed. Monitoring the use of medicines and medical devices after they have been approved and put on the market, in ever greater numbers each year, is an important public health issue. Monitoring each product systematically, by hand and on a case-by-case basis, does not seem to be a sustainable solution. Without completely replacing humans, AI can help to sort signals to determine which should be followed up on. Nevertheless, more recent events have shown us that currently, and probably for a long time yet, the best early warnings are based on the discernment of humans who are in the right place at the right time. It was not data-monitoring signals that raised the alarm about the COVID-19 pandemic, but a human doctor.

Learning Healthcare Systems (LHSs)

At the healthcare system level, for example, when there are one or more healthcare establishments involved, AI can be used to create a learning healthcare system, among other things [23]. This concept was developed in 2007 and has evolved since, with a focus on continuously improving the quality, organization, and suitability of care provided by a healthcare establishment. It involves collecting data about medical and paramedical practices (practice-based evidence), notably by using and processing electronic health records and utilizing flexible and continuous search abilities. Because an existing data warehouse or data pool is needed for an LHS to function, we see once again that AI can be used at various points throughout the system: from the collection of data to its specialized processing, federation, and

formatting. The AIs used in an LHS can vary in their autonomy: from very high (total) autonomy in the data extraction layers to much less autonomy (supervised) in the higher layers, where decisions are made. The prospects offered by LHSs require human involvement, both technical and organizational, with the promise of a system that is more reactive and less fixed. While the idea is interesting in principle, we lack sufficient information to accurately assess the realities and effectiveness of LHSs.

Risk and Insurance

Health is a special kind of economic good, as its capital can only ever diminish over time. Health costs are significant, and risk pooling and prevention mechanisms are often necessary, so that the individual whose risk materializes is not indebted for life or financially ruined because they have to pay the full costs themselves. We therefore have insurance, where you pay a premium based on the nature and likelihood of the risk you are insuring yourself against [24].

The notion of risk is central to public health, since it is used to determine primary, secondary, and tertiary prevention measures. Knowing all the relevant risk factors makes it possible to suggest changes in health behavior or to begin treatment before a pathology manifests itself [25].

The Notion of Risk in the Era of AI

Risk factors are not necessarily the cause of an illness, but they are indicators. Discovering a risk factor involves establishing a statistical association between an exposure (e.g., to tobacco) and a pathology (lung cancer) and then estimating the strength of association (relative risk). This is a familiar tool used in epidemiology [25]. AI's potential advantages lie in its ability to identify associations using new methods (machine learning), as well as in its ability to simultaneously account for a greater number of individual and contextual characteristics when calculating risk, and therefore risk intensity, for a specific person. Deep learning and Bayesian networks are particularly promising here. Bayesian networks make it

possible to calculate the conditional probability of developing a pathology given a set of characteristics [26]. Uncertainty and imprecision are taken into account in the network's design. Another advantage of Bayesian networks is that unlike most other AI algorithms, they provide a clear graphic model of the interconnections between characteristics. This helps to develop explanatory models for the health event in question: the model is already easier for human beings to understand.

Approaches Adopted from Marketing and PR: Segmentation and Targeting

The stratification of populations into different risk groups, traditionally following a gradient (of increasing risk), is similar to strategies used in marketing and PR. The idea is to segment a population based on certain characteristics, so that each segment can be targeted with qualitatively or quantitatively differentiated actions. A familiar example might be the strategies used by Facebook, Google, and Amazon, which have a lot to teach us about different techniques that could be used in public health to identify more relevant and more specific population strata. This approach has some similarities with individualization, since more characteristics can be taken into account, refining the way people are characterized by dividing them into smaller, and therefore more numerous, groups than in traditional approaches. For example, Amazon customers are divided into 200,000 different groups, and each customer may move from one group to another based on their browsing and purchase history. Each group is associated with a set of likely recommendations: if you belong to a certain group, because you "resemble" the other people in that group, it is likely that a purchase made by someone else in the group may be relevant to you, even though you did not make the purchase yourself. The stratification, segmentation, and targeting strategies used in marketing and PR are no longer unheard-of in public health. In fact, efforts to make preventative messages and information campaigns more effective already take their target audiences and communication methods into account.

Insurance and collective and individual risks can be integrated into the objectives of precision

public health. For example, precision public health may help to reduce health inequalities, including social and economic health inequalities, bringing a new dimension to the old issue of proportionate universalism [27]: should we help people in proportion to their health needs, with the goal of attaining a desired average level of population health? Or should the same assistance be provided to all, no matter their needs or socio-economic situation, in the interests of equality? In other words, because AI can help to identify risk factors that tend to be more individualized and more actionable than previous methods thanks to the use of digital technology and connected objects, the debate between equality and equity has returned to the fore.

Organization and Governance

Capturing Data for Research and Governance

As we have already said, one of the unique features of public health is that it operates at the intersection of health and policy and of research and medical practice. In fact, data and AI can be used to pursue two substantially different goals. A recent example that serves to illustrate how data and AI can be used across the spectrum between research and political decision-making is the COVID-19 pandemic and the role of case tracking in different public health policies aimed at controlling the epidemic at the national level [28]. In a highly aggregated form and initially intended for research purposes, mobile network operators made phone data available (e.g., in Italy and France), which painted a better picture of the changes in population movement before and after the implementation of lockdown measures. This information, initially used to calibrate models to track the geographical and populational spread of the virus (most models are “compartmental,” meaning they represent fragments of the population, with no need for individual data [29]), was actually used to inform policy decisions based on how effective restrictive measures had been and how well they had been followed. At the other end of the spectrum, smartphones could be

used as passive or active data collection tools (users enter the data themselves, or the phone passes on certain information, such as geolocation data), as information and communication tools (alerts when there has been suspected contact with a confirmed case, instructions or orders to go into quarantine, etc.), as well as a means for monitoring individuals. In the last instance, the smartphone can be seen as a digital passport, displaying a color code calculated by an AI whose criteria remain unknown and which decides whether or not to allow someone to enter a certain area (as in China [30]). The transfer of data between different actors, the police, airports, hospitals, health ministries, and so on, has also helped to drive individualized actions, for example, when a certain person was identified to be breaking their quarantine [31]. The police could then be informed and take action (as in South Korea and Singapore). More generally, over the last few years, we have seen the rise of personal scores (social credit score in China), calculated based on algorithms whose code remains secret [32]. The consequences of these scores are all too real, as they may bar citizens from accessing public resources, including healthcare. This is one of the downsides of AI: like most systems based on data federation, the mechanisms behind these scores still require human intervention.

New Public Health Actors: The Role of Platforms and the Private Sector

Digital technology is at the heart of AI’s contribution to public health. Digital technology has produced several giants of the private sector – the GAFAM (Google, Apple, Facebook, Amazon, and Microsoft) – who control as many resources and wield perhaps as much influence as most countries. These companies concentrate material resources and dominate the use of digital technologies by offering apparently free services that have become an essential part of daily life, allowing the companies to capture data through these tools and services. Their services have also been supported by the growth of AI: mostly for segmenting populations and targeting advertisements and marketing campaigns, but also more generally for refining recommendation systems

[33]. While health data in most countries is protected under a special legal status, a person's health status or diagnosis can often be better determined by capturing and cross-referencing data that are not initially related to health, but that are diverse and massive enough for the AI to interpret, than by using their actual health data. A significant portion of a person's health risk is based on their behaviors. The rise of open data has made it possible to cross-reference the private individual data captured by these companies with contextual and environmental data, for example, about exposures. No matter how you approach the issue, it is clear that these platforms are often best positioned to estimate risk, to open communication channels, and to personalize actions taken, since, beyond their digital and AI resources, these companies also possess incredible and very material logistical resources: data centers, warehouses, customer/vendor contact platforms, distribution networks, transportation networks, and so on.

Future Challenges

Many of the current and future challenges of AI in public health are not limited to this field alone. Larger issues remain with the acceptance of data capture and coding, with citizens' and patients' digital literacy [34], with biased algorithms whose faults hinder the further use of AI [35], with the digital divide, and finally with building algorithms that are powerful enough and appropriate for the field of health. There are also some challenges specific to the field of public health. For example, what trade-offs should be made when setting algorithms' optimization criteria, which need to cover both individual and collective health? And how much overlap should there be between health and policy?

Challenges Shared with Other Health Fields

AI will only become a part of public health if certain conditions are met: the relevant information systems would need to be interoperable and truly accessible [36], and there would need to be

support for capturing data and for processing and documenting this data. It will be necessary to win the acceptance and approval of citizens and patients, but also of healthcare professionals and decision-makers, for algorithms and the recommendations, information, etc., they produce, regardless of the quality of the algorithms developed [37, 38]. History has shown that the adoption of new digital technologies is not directly related to the quality or reliability of the information they provide, but rather to the user experience: Is it pleasant? Easy to use? Does it meet some need that the user has felt or previously expressed?

Biases are an important problem [35]. Algorithms are constructed based on data. In the frequent case of the reuse of data – that is, the use of data that was not collected specifically for building the algorithm – various biases may be found, including those of the designers themselves. Foremost among these are selection biases, since the database may not contain all of the relevant information. Instances of biases based on gender, ethnicity, and geographical origin have been the best documented so far: the risk in applying an algorithm developed from health data about white men to women of African origin is that it may produce results that are irrelevant and possibly even dangerous. This bias is already present in clinical research, but it is exacerbated by the use of AI, especially in public health, which covers many diverse populations. The hidden danger of highly inclusive AI in public health is that it would tend toward a totalizing representation and utilization. This risk stems from the principles of public health itself, which could be used by the state to try to organize every aspect of individual life in order to protect public health (or security, etc.). More generally, problems of health and digital literacy still need to be considered [34]. Not everyone has the same resources for engaging with the digital world, and the use of AI in public health could widen existing inequalities, or create new ones. Beyond simple digital literacy, access to digital hardware is also a real problem. During the COVID-19 pandemic, the population most at risk of serious illness and death is the elderly. In most countries, however, this population also has the lowest level of digital literacy and is the least likely to own a smartphone. This means that strategies to control

the epidemic based on smartphones and AI might not be the most suitable.

More Specific Challenges

It is often difficult to predict whether new phenomena in healthcare will be useful, since there are so many standards to be met. While a marketing department might be happy with a conversion rate (how many people see an advertisement vs how many people make a purchase) of between 10 and 40 percent, these levels of absolute effectiveness and relative gain are well below what is necessary in healthcare. For certain very specific and well-defined medical fields and tasks, such as anatomical pathology, AI performance may be excellent, especially when the AI is designed to support humans. But this is not true across the board, and most specialities require some proof that it is worth adding another dimension of complexity, namely, the social and populational dimension.

The tension between individual preference and social preference is present from the very first stage of algorithm design, even if it is rarely discussed. Essentially, algorithm training follows an optimization function. Most of the time, the algorithm looks for the best possible optimum for a specific criterion. In terms of public health, what exactly should be optimized, potentially to the detriment of other criteria? Individual health? Collective health? Cost-effectiveness? A technical debate on multi-criteria optimization and a social debate are both needed on this issue.

Of course, all of the challenges that AI normally faces in terms of ethics, societal issues, and legal issues take on a special importance here, since it is not only a question of individual health, but also of health policy. Whether the line between the public and the private in these matters should be sharply drawn or somewhat blurred remains an important question. Until now, the private sector has dominated the industrial aspects of health: producing medicines and medical devices. This sector is subject to heavy regulation at the national and international levels. It is therefore difficult for AI and digital technology to gain a foothold, since incredibly powerful private actors are already present upstream. These private

actors may seek to outflank newcomers by influencing new regulations.

Conclusion and Outlook

Artificial intelligence is used in public health at at least two points in the process of turning data into decision-making: upstream, in the capture, federation, and formatting of data, and downstream, in supporting decision-making. Until now, AI has developed rather quietly, focusing intently on data and only drawing the attention of those with a technical interest. If AI is to be used in decision-making, including any important contributions to the creation and implementation of health policy, certain issues must be resolved: namely, the dual tension between optimizing individual criteria or collective criteria and between public prerogatives and private ambitions.

Cross-References

- [Artificial Intelligence in Epidemiology](#)

References

1. Wade DT, Halligan PW. The biopsychosocial model of illness: a model whose time has come. *Clin Rehabil.* 2017;31(8):995–1004. <https://doi.org/10.1177/0269215517709890>.
2. Fassin D. Santé Publique. In: Lecourt D, editor. *Dictionnaire de la pensée médicale*. Paris: PUF; 2004. p. 1014–8.
3. Dubé E, Laberge C, Guay M, Bramadat P, Roy R, Bettinger J. Vaccine hesitancy: an overview. *Hum Vaccin Immunother.* 2013;9(8):1763–73. <https://doi.org/10.4161/hv.24657>.
4. Ward JK, Cafiero F, Fretigny R, Colgrove J, Seror V. France’s citizen consultation on vaccination and the challenges of participatory democracy in health. *Soc Sci Med.* 2019;220:73–80. <https://doi.org/10.1016/j.soscimed.2018.10.032>.
5. Lee PR. The future of social medicine. *J Urban Health.* 1999;76(2):229–36. <https://doi.org/10.1007/BF02344678>.
6. Kawachi I, Subramanian SV. Social epidemiology for the 21st century. *Soc Sci Med.* 2018;196:240–5. <https://doi.org/10.1016/j.soscimed.2017.10.034>.
7. Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine. *Nat Rev Clin Oncol.* 2011;8:184–7.

8. Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Med Inform Decis Mak.* 2019;19(1):211. <https://doi.org/10.1186/s12911-019-0918-5>.
9. Lin E, Lin CH, Lane HY. Precision psychiatry applications with pharmacogenomics: artificial intelligence and machine learning approaches. *Int J Mol Sci.* 2020;21(3):969. <https://doi.org/10.3390/ijms21030969>.
10. Lillie EO, Patay B, Diamant J, et al. The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? *Pers Med.* 2011;8(2):161–73. <https://doi.org/10.2217/pme.11.7>.
11. Seeking precision in public health. *Nat Med.* 2019;25(8):1177. <https://doi.org/10.1038/s41591-019-0556-6>
12. Horton R. Offline: in defence of precision public health. *Lancet.* 2018;392(10157):1504. [https://doi.org/10.1016/S0140-6736\(18\)32741-7](https://doi.org/10.1016/S0140-6736(18)32741-7).
13. Godlee F. Evidence based medicine: flawed system but still the best we've got. *BMJ.* 2014;348:g440.
14. Kiran T. Toward evidence-based policy. *CMAJ.* 2016;188(15):1065–6. <https://doi.org/10.1503/cmaj.160692>.
15. Latour B, Woolgar S. Laboratory life: the social construction of scientific facts. Los Angeles: Sage; 1979.
16. Anderson C. The end of theory: the data deluge makes the scientific method obsolete. *Wired,* 2008. <https://www.wired.com/2008/06/pb-theory>
17. Rice MJ, Stalling J, Monasterio A. Psychiatric-mental health nursing: data-driven policy platform for a psychiatric mental health care workforce. *J Am Psychiatr Nurses Assoc.* 2019;25(1):27–37. <https://doi.org/10.1177/1078390318808368>.
18. Kamel Boulos MN, Peng G, VoPham T. An overview of GeoAI applications in health and healthcare. *Int J Health Geogr.* 2019;18(1):7. <https://doi.org/10.1186/s12942-019-0171-2>.
19. Huang P, MacKinlay A, Yepes AJ. Syndromic surveillance using generic medical entities on Twitter. In: Proceedings of Australasian language technology association workshop, 2016. p. 35–44.
20. Hamon T, Gagnayre R. Improving knowledge of patient skills thanks to automatic analysis of online discussions. *Patient Educ Couns.* 2013;92(2):197–204. <https://doi.org/10.1016/j.pec.2013.05.012>.
21. Chiolero A, Buckeridge D. Glossary for public health surveillance in the age of data science. *J Epidemiol Community Health.* 2020;74:612–6.
22. Kandula S, Shaman J. Reappraising the utility of Google Flu Trends. *PLoS Comput Biol.* 2019;15(8):e1007258. <https://doi.org/10.1371/journal.pcbi.1007258>.
23. Wongvibulsin S, Zeger SL. Enabling individualised health in learning healthcare systems. *BMJ Evid Based Med.* 2020;25(4):125–9. <https://doi.org/10.1136/bmjebm-2019-111190>.
24. Ho CWL, Ali J, Caals K. Ensuring trustworthy use of artificial intelligence and big data analytics in health insurance. *Bull World Health Organ.* 2020;98(4):263–9. <https://doi.org/10.2471/BLT.19.234732>.
25. Cole SR, Hudgens MG, Brookhart MA, Westreich D. Risk. *Am J Epidemiol.* 2015;181:246–50. <https://doi.org/10.1093/aje/kwv001>.
26. Lefèvre T, Lepresle A, Chariot P. Detangling complex relationships in forensic data: principles and use of causal networks and their application to clinical forensic science. *Int J Legal Med.* 2015;129(5):1163–72. <https://doi.org/10.1007/s00414-015-1164-8>.
27. Marmot M. Fair society, healthy lives: the Marmot Review: strategic review of health inequalities in England post-2010. 2010. ISBN 9780956487001.
28. Bengio Y. <https://yoshuabengio.org/fr/2020/03/25/depistage-pair-a-pair-de-la-covid-19-base-sur-lia/>
29. Kröger M, Schlickeiser R. Analytical solution of the SIR-model for the temporal evolution of epidemics. Part A: time-independent reproduction factor. *J Phys A.* 2020. <https://doi.org/10.1088/1751-8121/abc65d>.
30. Mozour P, Zhong R, Krolik A. In coronavirus fight, China gives citizens a color code, with red flags. *The New York Times,* 2020. <https://www.nytimes.com/2020/03/01/business/china-coronavirus-surveillance.html>
31. Lee Y. Taiwan's new 'electronic fence' for quarantines leads wave of virus monitoring. *Reuters,* 2020. <https://www.reuters.com/article/us-health-coronavirus-taiwan-surveillanc/taiwans-new-electronic-fence-for-quarantines-leads-wave-of-virus-monitoring-idUSKBN2170SK>
32. Bach J. The red and the black: China's social credit experiment as a total test environment. *Br J Sociol.* 2020;71(3):489–502. <https://doi.org/10.1111/1468-4446.12748>.
33. Tran TNT, Felfernig A, Trattner C, et al. Recommender systems in the healthcare domain: state-of-the-art and research issues. *J Intell Inf Syst.* 2020. <https://doi.org/10.1007/s10844-020-00633-6>.
34. Manganello J, Gerstner G, Pergolino K, Graham Y, Falisi A, Strogatz D. The relationship of health literacy with use of digital technology for health information: implications for public health practice. *J Public Health Manag Pract.* 2017;23(4):380–7. <https://doi.org/10.1097/PHH.0000000000000366>.
35. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 2019;366(6464):447–53. <https://doi.org/10.1126/science.aax2342>.
36. Unberath P, Prokosch HU, Gründner J, Erpenbeck M, Maier C, Christoph J. EHR-independent predictive decision support architecture based on OMOP. *Appl Clin Inform.* 2020;11(3):399–404. <https://doi.org/10.1055/s-0040-1710393>.
37. Chiang J, Kumar A, Morales D, Saini D, Hom J, Shieh L, Musen M, Goldstein MK, Chen JH. Physician usage and acceptance of a machine learning recommender system for simulated clinical order entry. *AMIA Jt Summits Transl Sci Proc.* 2020;2020:89–97.
38. Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak.* 2020;20(1):310. <https://doi.org/10.1186/s12911-020-01332-6>.



AIM and Business Models of Healthcare

43

Edward Christopher Dee, Ryan Carl Yu, Leo Anthony Celi, and
Umbereen Sultana Nehal

Contents

Introduction	604
The Business Perspective: Product Development and Sales	604
The Consumer Perspective: Purchaser, End User, and Patient	605
Myth of Generalizability	606
Proposed Consideration 1: Co-creation	607
Proposed Consideration 2: Multi-Stakeholder Engagement	607
Proposed Consideration 3: Metrics for Defining Value Delivered	607
Ethics, Law, and Policy	608

E. C. Dee
Harvard Medical School, Boston, MA, USA
e-mail: deee1@mskcc.org

R. C. Yu
Harvard Business School, Cambridge, MA, USA
e-mail: ryu@mba2021.hbs.edu

L. A. Celi (✉)
Institute for Medical Engineering and Science, Massachusetts
Institute of Technology, Cambridge, MA, USA
Department of Biostatistics, Harvard T.H. Chan School of
Public Health, Boston, MA, USA

Laboratory for Computational Physiology, Massachusetts
Institute of Technology, Cambridge, MA, USA

Division of Pulmonary, Critical Care and Sleep Medicine,
Beth Israel Deaconess Medical Center, Boston, MA, USA
e-mail: lceli@mit.edu

U. S. Nehal
MIT Sloan School of Management, Cambridge, MA, USA
University of Massachusetts Medical School, Worcester,
MA, USA
e-mail: usnehal@mit.edu

Conclusion	609
References	609

Abstract

Artificial intelligence (AI) and machine learning in healthcare are growing at an unprecedented rate. Myriad uses of medical AI, ranging from tumor identification on imaging to workforce management, make use of a wealth of available healthcare data. These models are becoming increasingly commercially available. However, much of the utility of medical AI depends on the quality of the data models trained on, and critically, the contexts and biases within which these models are created. In this chapter, we first describe a business-informed framework that influences product development and commercialization of these technologies. We describe the consumer side that includes purchasers, end users, and patients. Subsequently, we underscore the pitfalls of the assumption that models trained in one context can be applied to another, that is, the myth of generalizability. We propose solutions to these problems and describe the importance of co-creation and multi-stakeholder engagement in designing medical AI. We highlight the need to define value metrics that consider equity and the mitigation of healthcare disparities. Lastly, we draw attention to open ethical, legal, and policy questions that must be answered as the role of AI in medicine progresses and grows.

Introduction

Paul Gauguin's magnum opus "Where Do We Come From? What Are We? Where Are We Going?" appears to outline the human life cycle. Read from right to left, the painting poses existential questions about what happens beyond. It is an apt analogy for current times of rapid transformation within healthcare, an industry that accounts for about one-fifth of the US economy. Specifically, artificial intelligence (AI) and machine learning (ML) in healthcare are projected

to grow from \$4.9 billion in 2020 to \$45.2 billion by 2026 [1]. Examples of medical AI abound; recent studies make use of AI to identify tumors on imaging [2, 3], direct aspects of clinical evaluation [4], and workforce management [5].

Evidence-based and data-driven medicine is considered the gold standard of high-quality, safe healthcare. Additionally, the healthcare industry is increasingly shifting toward value-based care models of payment [6]. This creates incentives to rely increasingly on big data [7]. A commonly used phrase applied to big data analytics is "garbage in, garbage out." If we are to deliver on the promise of AI with effective implementation, that requires specific frameworks and intentional harm mitigation to protect from unintended consequences.

In this chapter, we present a business perspective of medical AI and identify limits of ready-to-use off-the-shelf medical AI algorithms that are increasingly common. We offer an alternative approach with which to create and incorporate AI-based medical tools with strategies for high value purchasing decisions and implementation. Lastly, we will discuss the key considerations around AI-enabled healthcare technologies that incorporate quality, value, accuracy, effectiveness, and equity.

The Business Perspective: Product Development and Sales

It is important to consider the process of product development from the business and innovation point of view. To "know your customer," in the complexity of healthcare, we must define who "the customer" is. We must differentiate the end user from the economic buyer from the patient or consumer of care. Specifically, the former may be a clinician who is using hospital-purchased AI to diagnose and treat patients. The economic buyer would be a director or c-suite executive making the purchasing decision at an institutional level. In contrast, the healthcare consumer is the patient

whose health outcomes may depend on the quality of AI used but often has no say or knowledge of AI use or purchase. Thus, the target customer for a company developing AI for healthcare may be removed from the experience of the end user or the patient. This may result in misaligned incentives or gaps.

Also, whereas many clinicians consider healthcare to be a social good and ideally do not differentiate between one patient and another, for product developers and business, there is often a deliberate approach to market segmentation. That means intentionally selecting a particular segment of the market based on disease type, geography, population density, ability or willingness to pay, cost of customer acquisition (CoCA), etc. As medical AI falls into business-to-business (B2B) category, innovation firms seek customers with efficient sales processes. Some customers may have different size and complexity of the decision-making unit (DMU), which in turn affects CoCA.

These factors can influence initial access to healthcare technology as well as which data are used for initial algorithm development. It is plausible that the least discerning customer with fewer internal safeguards is the first purchaser. On the other hand, it may be that the savviest early adopter may influence the product by being part of the beachhead market. Later adopters, who believe by waiting they are getting a “proven” product, are actually purchasing something not necessarily equally effective in their own population.

Once a product is developed, its use may be broadened to populations and settings for which it was not originally designed or in which it was not tested. The “myth of generalizability” – the oftentimes incorrect assumption that conclusions drawn from one population apply to another – affects healthcare AI as it has affected much more established realms of healthcare [8–10]. Although many medical journals require that articles presenting healthcare AI validate results on external cohorts, this may not be a consideration for all potential purchasers and users of healthcare AI [11, 12]. Many early evaluations of medical AI have demonstrated that strong performance at a single site may not translate to strong performance at many other sites; there seems to be an inherent trade-off between improving systems locally and

creating generalizable systems [9, 13–15]. Therefore, marketing claims on accuracy, effectiveness, or predicted outcomes using data from other customers must be evaluated with a healthy degree of skepticism.

The Consumer Perspective: Purchaser, End User, and Patient

A key consideration is that accuracy of AI – much like the application of randomized controlled trials [16–18] – is bound by space and time. In particular, as a “learning” system, medical AI requires the relevant and most matched dataset not only for training but for continuous updates and ongoing improvements in order to serve the healthcare AI purchaser at a systems level, the end user of healthcare AI, as well as the patient or healthcare consumer. With regard to systems-level purchasers, examples include clinical systems, payers, accountable care organizations, or government entities. As these systems may have high complexity or large DMUs, they may be the last to be approached for sales and/or slowest to adopt technology.

Importantly, the populations that medical AI is meant to learn from – and the populations medical AI is meant to serve – are dynamic and diverse. In contrast, many algorithms are created and sold as “ready-made,” that is, trained and proprietary algorithms that are meant to be used as-is by the user. Furthermore, algorithms based on a finite set of data may not be able to adapt to different populations not captured by the data on which the algorithms are trained. If many of these algorithms do not correct for biases that are systemic in the data from which the algorithms are made to learn, then seemingly “novel” technology may perpetuate biases of the past.

An example of this would be that AI designed to detect radiologic lesions that is trained and tested in one ethnic or geographic population may not have the correct standards or benchmarks when applied to another population [19]. This is especially salient when considering existing racial and gender disparities that exist in data used in clinical trials that inform our current evidence-based medicine standards [20]. Data used for

current guidelines comes from studies with primarily white and male clinical subjects [20–22]. The majority of dermatology images, an area ripe for AI, use white subjects [23]. Facial recognition technology has been found to be more commonly inaccurate if users are female and/or black [24]. As we are starting with existing biases in our data, then we must be careful and resolute that AI design and implementation is suited to deliver quality outcomes across diverse populations. Another area of healthcare disparity is rural health. Use of AI has potential to close gaps especially on improving resource allocation, promoting supply chain integrity, and addressing workforce shortages [25]. This requires further study to evaluate.

Myth of Generalizability

Medical AI relies on the computer's ability to synthesize certain kinds of data almost infinitely faster than a human can and on the computer's ability to store vast amounts of information in its memory. In teaching computers to "see" (computer vision), "listen" (natural language processing), or "think" (CNNs), medical AI models are often built to sense patterns from quantities of information that are much larger than humans can. However, the assumptions made by AI may be predicated on biases inherent to the teams developing these tools [26].

Two examples are illustrative of the importance of multicenter data with local validation. The first, in ophthalmology [27], provides evidence for the validation of an AI-based model when comparing data from two ophthalmologic institutions in India within a field where there is known specialist variation. Another study of AI for chest X-rays in the United States demonstrates that even within a single country, models may not generalize across institutions.

The study by Gulshan et al. sought to validate the performance of an automated diabetic retinopathy (DR) diagnosis system across two sites in India. They found that, using data prospectively collected for two ophthalmologic institutions in India, an automated algorithm identified referable DR with performance equal to or exceeding the

retinal specialists and trained graders for a discrete outcome [27]. While the international retinal grading system is on a more granular five-point scale, it is common for clinical decision-making to focus on what is actionable or not. The gold standard used was labels agreed upon by a panel of retinal specialists. The need for a consensus DR classification has been noted in multiple studies to account for variation in specialist readings [28, 29]. Gulshan et al. caution that the use of different types of cameras may also impact the generalizability of their algorithm. Subsequent research on AI-powered DR evaluation by Google Health in Thailand across multiple centers found that clinical criteria alone were insufficient without data on socio-environmental factors [30]. For the healthcare administrator or government official seeking to apply this at the systems level, the takeaways are that such AI may be used as a screening tool, does not replace human specialists, and must take into account local context and data.

Another example is of Zech et al. [31], in which the authors employed convolutional neural networks (CNNs) to analyze medical imaging and generate computer-aided diagnoses. They assessed how well CNNs generalized across 3 hospital systems for a simulated pneumonia screening task, making use of a cohort of 158,323 chest radiographs drawn from 3 different institutions in the United States. Models trained on one institution's data did not always perform well when validated on another institution's data. For example, they found that the highest internal performance was achieved by combining training and test data from two of the three institutions (AUC 0.931, 95% CI 0.927–0.936); however, the model demonstrated significantly lower external performance at the third institution (AUC 0.815, 95% CI 0.745–0.885, $P = 0.001$). The authors advise: "Given the significant interest in using deep learning to analyze radiological imaging, our findings should give pause to those considering rapid deployment of such systems without first assessing their performance in a variety of real-world clinical settings." [31]

Taken together, these examples underscore the potential pitfalls in generalizability. Healthcare executives who seek to replace costly specialists

with medical AI are advised extreme caution; the evidence does not readily support that approach for consistent quality in patient-level or population-level outcomes. At the same time, in low resource settings with workforce shortages, as is found in rural health and global health, AI for screening may assist with prioritization or resource allocation decisions. Understanding strengths and limitations of any technology solution allows effective implementation.

Given these challenges, we offer the following approaches to increase value and mitigate risks:

Proposed Consideration 1: Co-creation

A best practice in product development involves co-creation, customization to the consumer, use of consumer data [32], short product life cycle, and agile approach with continuous releases. This is a change from traditional product development that employed a “one-size-fits-all approach” to development of a minimum viable product (MVP), longer timelines, and longer product life cycle. We believe it is critical for healthcare AI to utilize these industry best practices of ensuring that technology used is trained on local data from the AI purchaser rather than from a different dataset. The very nature of AI and ML allows the product to be customized to the customer if the ongoing learning is on customer-specific data.

Despite variation in context and patient population affecting the systemic and patient-level generalizability of many ready-made medical AI technologies [9, 13–15], the continuous learning nature of AI and ML could offer an advantage over standard use of “one-size-fits-all” clinical practice guidelines. By employing local data from which to draw inferences, clinician decisions can be more matched to the local population [33–36].

Diversification of training data is an important first step: as with clinical trial data, the more closely the study cohort mimics the clinical context in which AI solutions are being employed, the greater the accuracy [37, 38]. There is a false assumption that those informing AI development must be well versed in technology or innovation. In fact, local practitioners provide the requisite

insights into the particular clinical questions that AI solutions may help solve.

Proposed Consideration 2: Multi-Stakeholder Engagement

The public sector and nonprofits, whose mission is often to serve the needs of a wide array of groups, are a useful source of inclusive best practices for multi-stakeholder engagement throughout the stages of product development. For example, the Patient-Centered Outcomes Research Institute (PCORI) has developed a robust rubric used for patient-centered study design and widely accepted framework employed in the context of healthcare research and innovation that serves both large systems and smaller groups that face health disparities [39, 40]. Combining medical patient-engaged research design with innovative human-centered design is recommended for businesses in medical AI product development [41].

An additional benefit of this approach is staying abreast of the evolving legal and policy landscape discussed later in the chapter. While business approaches sometimes can limit information sharing, in fact, open channels of communication can allow access to the most up-to-date knowledge through collaborators and engaged stakeholders. Engaged stakeholders can inform strategic growth, mitigate risks, and identify future avenues of dissemination.

Proposed Consideration 3: Metrics for Defining Value Delivered

Increasing penetration of value-based care models requires careful definition of quality. Value metrics for medications and machines exist, ranging from changes in biomarker levels, survival, and cost-effectiveness; however, these may be replete with biases or may not truly benefit the patient or end user. Biomarkers may not account for patient diversity. Survival outcomes may not account for adverse effects. Cost-effectiveness may not capture the extent of a patient’s experience. Fields are continuing to modify their metrics to adopt these dynamic values. For example, the oncology

community is growing in its focus on patient-centered outcomes rather than survival alone in its evaluation of treatment options [42]. Many fields are recognizing financial toxicity as another adverse effect of treatments [43, 44]. Importantly, for medical AI, we are at the position to define metrics for success with which we measure progress.

Reducing disparities has already been defined as value-based care quality metrics [45–47]. Some large payers institute payment withholdings if quality targets are not met [48]. We propose medical AI should be evaluated by matching metrics to those clinical systems already held accountable for payment. For example, it has been shown that women treated by female doctors have better myocardial infarction outcomes [49] and Black infants treated by Black pediatricians may have lower mortality [50, 51]. Medical AI trained to help predict risks and determine treatment should aim to close these gaps [52, 53].

It is critical to understand that conclusions from data of certain systems may imply causation related to individual factors that are, in fact, systems issues, as with the use of big data in health insurance [54]. Healthcare purchasers who do not ask discerning questions may find poorly validated medical AI harms both their patients and their bottom line. Business executives who are thoughtful can differentiate their medical AI products by comprehensive quality metrics. These cautions we offer highlight an opportunity to improve and represent a challenge to innovators.

Ethics, Law, and Policy

In forging the way forward, we must pause and ask “Where are we going?” We need to ensure that we aim to do more than convert existing data into technological solutions; we must aim to improve both data sources and outcomes. Within the field of ethics and bioethics for medical AI are many potential approaches defined by culture, values, philosophy, political affiliation, and national interest, all of which require careful thought.

Of note, Gauguin’s “Where Do We Come From? What Are We? Where Are We Going?” was painted at a time of personal crisis. It depicts the artist’s transition from centering on Judeo-Christian references to subjects interwoven with Tahitian mythology [55]. Similarly, in AI, we must be thoughtful in what use as a reference value or school of thought. The concept of “crisis” is particularly relevant during the COVID-19 pandemic, and simultaneously, at a time during which the American Medical Association labeled structural racism as a “public health threat” [56].

Additionally, there is a dynamic tension between industry controls for brand management and academic intellectual freedom. How these interfaces are handled can have significant consequences for trust in companies and industry for thought leadership, as was the case for Google’s handling of research ethics and equity [57]. Continued collaboration across sectors is essential to promote innovation and broad dissemination.

Legal liability of AI use in healthcare is an evolving field – who owns the risk when errors occur? From the example of electronic health record, even system failures can result in physician liability [58]. This is especially complicated in an industry where, per one study of US healthcare, error was already the third leading cause of death [59]. Of note, the critics of that study consider the large database analysis inadequate, which underscores the uncertainty inherent to use of big data to define baselines and benchmarks. As there will never be a single right answer, we advise developers and purchasers of medical AI to seek legal counsel with regard to enterprise risks. It is important to establish a clear rationale and documentation for potential future legal challenges.

Further, international law, particularly from the European Union [60], is shaping the opportunities and limits of this industry, with ongoing debates related to privacy, data sharing, competition, and even national security. The potential for a new technology “Cold War” has been suggested [61]. Specifically, there is a call for technology diplomacy, that is, ongoing efforts to bolster beneficence and equity in the dissemination and application of medical AI on a global scale

[62]. Evolving policy at institutional, local, state, national, and international level must be continually referenced and adapted to as we wrestle with complex and often contentious ethical dilemmas of medical AI.

Conclusion

In this chapter, we have asked the reader to consider how to ensure medical AI delivers value to customers. We offer cautious optimism for a technologically advanced future that serves both human needs and business objectives of revenue generation, market share, and return on investment. If innovation is opportunistic, fails to examine structural inequity, or does not engage stakeholders, there is the danger of merely encoding and perpetuating the past. On the other hand, with intentional human-centered design that accounts for the range of human variation across contexts, there is boundless opportunity. Therefore, there is urgent and deliberate need to be conscientious in our development and adoption of medical AI. In doing so, we may not need to ask “Where are we going?” but instead define and create the way forward.

References

- [62]. Evolving policy at institutional, local, state, national, and international level must be continually referenced and adapted to as we wrestle with complex and often contentious ethical dilemmas of medical AI.
-
- (SHIELD-RT): a prospective randomized study of machine learning-directed clinical evaluations during radiation and chemoradiation. *J Clin Oncol*. Published online 2020. <https://doi.org/10.1200/JCO.20.01688>.
5. Spatharou A, Hieronimus S, Jenkins J. Transforming healthcare with AI: the impact on the workforce and organizations. McKinsey & Company.
 6. Berwick DM. Elusive waste: the Fermi Paradox in US health care. *JAMA – J Am Med Assoc*. Published online 2019. <https://doi.org/10.1001/jama.2019.14610>.
 7. Schneweiss S. Learning from big health care data. *N Engl J Med*. Published online 2014. <https://doi.org/10.1056/nejm1401111>.
 8. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ*. Published online 2020. <https://doi.org/10.1136/bmj.m1328>.
 9. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health*. Published online 2020. [https://doi.org/10.1016/S2589-7500\(20\)30186-2](https://doi.org/10.1016/S2589-7500(20)30186-2).
 10. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res*. Published online 2016. <https://doi.org/10.2196/jmir.5870>.
 11. Bluemke DA, Moy L, Bredella MA, et al. Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers—from the Radiology Editorial Board. *Radiology*. Published online 2020. <https://doi.org/10.1148/radiol.2019192515>.
 12. Leisman DE, Harhay MO, Lederer DJ, et al. Development and reporting of prediction models. *Crit Care Med*. Published online 2020. <https://doi.org/10.1097/CCM.0000000000004246>.
 13. Bedoya AD, Clement ME, Phelan M, Steorts RC, O'Brien C, Goldstein BA. Minimal impact of implemented early warning score and best practice alert for patient deterioration. *Crit Care Med*. Published online 2019. <https://doi.org/10.1097/CCM.0000000000003439>.
 14. Downey CL, Tahir W, Randell R, Brown JM, Jayne DG. Strengths and limitations of early warning scores: a systematic review and narrative synthesis. *Int J Nurs Stud*. Published online 2017. <https://doi.org/10.1016/j.ijnurstu.2017.09.003>.
 15. Gerry S, Bonnici T, Birks J, et al. Early warning scores for detecting deterioration in adult hospital patients: systematic review and critical appraisal of methodology. *BMJ*. Published online 2020. <https://doi.org/10.1136/bmj.m1501>.
 16. Rothwell PM. Factors that can affect the external validity of randomised controlled trials. *PLoS Clin Trials*. Published online 2006. <https://doi.org/10.1371/journal.pctr.0010009>.
 17. Rothwell PM. External validity of randomised controlled trials: “to whom do the results of this trial apply?” *Lancet*. Published online 2005. [https://doi.org/10.1016/S0140-6736\(04\)17670-8](https://doi.org/10.1016/S0140-6736(04)17670-8).

18. Jüni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. *Br Med J*. Published online 2001. <https://doi.org/10.1136/bmj.323.7303.42>.
19. Geis JR, Brady A, Wu CC, et al. Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement. *Insights Imaging*. Published online 2019. <https://doi.org/10.1186/s13244-019-0785-8>.
20. Rencsok EM, Bazzi LA, McKay RR, et al. Diversity of enrollment in prostate cancer clinical trials: current status and future directions. *Cancer Epidemiol Biomarkers Prev*. Published online 2020. <https://doi.org/10.1158/1055-9965.EPI-19-1616>.
21. Murthy VH, Krumholz HM, Gross CP. Participation in cancer clinical trials: race-, sex-, and age-based disparities. *J Am Med Assoc*. Published online 2004. <https://doi.org/10.1001/jama.291.22.2720>.
22. King TE. Racial disparities in clinical trials. *N Engl J Med*. Published online 2002. <https://doi.org/10.1056/nejm200205023461812>.
23. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. *JAMA Dermatol*. Published online 2018. <https://doi.org/10.1001/jamadermatol.2018.2348>.
24. Buolamwini J. Gender shades: intersectional accuracy disparities in commercial gender classification supplementary materials. 2018.
25. Wahl B, Cossy-Gantner A, Germann S, ... Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? *BMJ Glob Heal*. Published online 2018.
26. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA – J Am Med Assoc*. Published online 2019. <https://doi.org/10.1001/jama.2019.18058>.
27. Gulshan V, Rajan RP, Widner K, et al. Performance of a deep-learning algorithm vs manual grading for detecting diabetic retinopathy in India. *JAMA Ophthalmol*. Published online 2019. <https://doi.org/10.1001/jamaophthalmol.2019.2004>.
28. Krause J, Gulshan V, Rahimy E, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*. Published online 2018. <https://doi.org/10.1016/j.ophtha.2018.01.034>.
29. Schaeckermann M, Hammel N, Terry M, et al. Remote tool-based adjudication for grading diabetic retinopathy. *Transl Vis Sci Technol*. Published online 2019. <https://doi.org/10.1167/tvst.8.6.40>.
30. Beede E, Baylor E, Hersch F, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: Conference on human factors in computing systems – proceedings. 2020. <https://doi.org/10.1145/3313831.3376718>.
31. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med*. Published online 2018. <https://doi.org/10.1371/journal.pmed.1002683>.
32. Geissbauer R, Wunderlin J, Schrauf S, et al. Digital Product Development 2025: agile, collaborative, AI driven and customer centric. PricewaterhouseCoopers GmbH Wirtschaftsprüfungsgesellschaft. Published 2019. <https://www.pwc.de/de/digitale-transformation/pwc-studie-digital-product-development-2025.pdf>
33. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med*. Published online 2018. <https://doi.org/10.1038/s41591-018-0213-5>.
34. Hampton JR. Evidence-based medicine, opinion-based medicine, and real-world medicine. *Perspect Biol Med*. Published online 2002. <https://doi.org/10.1353/pbm.2002.0070>.
35. Sherman RE, Anderson SA, Dal Pan GJ, et al. Real-world evidence – what is it and what can it tell us? *N Engl J Med*. Published online 2016. <https://doi.org/10.1056/NEJMsb1609216>.
36. Panch T, Pollard TJ, Mattie H, Lindemer E, Keane PA, Celi LA. “Yes, but will it work for my patients?” Driving clinically relevant research with benchmark datasets. *npj Digit Med*. Published online 2020. <https://doi.org/10.1038/s41746-020-0295-6>.
37. Deo RC. Machine learning in medicine. *Circulation*. Published online 2015. <https://doi.org/10.1161/CIRCULATIONAHA.115.001593>.
38. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf*. Published online 2019. <https://doi.org/10.1136/bmjqqs-2018-008370>.
39. Sheridan S, Schrandt S, Forsythe L, Hilliard TS, Paez KA. The PCORI engagement rubric: promising practices for partnering in research. *Ann Fam Med*. Published online 2017. <https://doi.org/10.1370/afm.2042>.
40. Patient-Centred Outcomes Research Institute. Engagement rubric for applicants. 2014. Published online 2016.
41. Boaz A, Hanney S, Borst R, O’Shea A, Kok M. How to engage stakeholders in research: design principles to support improvement. *Heal Res Policy Syst*. Published online 2018. <https://doi.org/10.1186/s12961-018-0337-6>.
42. Oliver A, Greenberg CC. Measuring outcomes in oncology treatment: the importance of patient-centered outcomes. *Surg Clin North Am*. Published online 2009. <https://doi.org/10.1016/j.suc.2008.09.015>.
43. Valero-Elizondo J, Khera R, Saxena A, et al. Financial hardship from medical bills among nonelderly U.S. adults with atherosclerotic cardiovascular disease. *J Am Coll Cardiol*. 2019;73(6):727–32. <https://doi.org/10.1016/j.jacc.2018.12.004>.
44. Knight TG, Deal AM, Dusetzina SB, et al. Financial toxicity in adults with cancer: adverse outcomes and noncompliance. *J Oncol Pract*. 2018;14(11):e665–73. <https://doi.org/10.1200/jop.18.00120>.

45. Thurman WA, Harrison T. Social context and value-based care: a capabilities approach for addressing health disparities. *Policy Polit Nurs Pract*. Published online 2017. <https://doi.org/10.1177/1527154417698145>.
46. Casalino LP, Elster A. Will pay-for-performance and quality reporting affect health care disparities? *Health Aff*. Published online 2007. <https://doi.org/10.1377/hlthaff.26.3.w405>.
47. Alberti PM, Bonham AC, Kirch DG. Making equity a value in value-based health care. *Acad Med*. Published online 2013. <https://doi.org/10.1097/ACM.0b013e3182a7f76f>.
48. Musser E. Measuring for equity: the medicaid quality network. *NCQA Blog*.
49. Greenwood BN, Carnahan S, Huang L. Patient–physician gender concordance and increased mortality among female heart attack patients. *Proc Natl Acad Sci U S A*. Published online 2018. <https://doi.org/10.1073/pnas.1800097115>.
50. Mahase E. Black babies are less likely to die when cared for by black doctors, US study finds. *BMJ*. Published online 2020. <https://doi.org/10.1136/bmj.m3315>.
51. Greenwood BN, Hardeman RR, Huang L, Sojourner A. Physician-patient racial concordance and disparities in birthing mortality for newborns. *Proc Natl Acad Sci U S A*. Published online 2020. <https://doi.org/10.1073/pnas.1913405117>.
52. Schuster A, Lange T, Backhaus SJ, et al. Artificial intelligence based fully automated myocardial function assessment for diagnostic and prognostic stratification following myocardial infarction. *J Am Coll Cardiol*. Published online 2020. [https://doi.org/10.1016/s0735-1097\(20\)32192-6](https://doi.org/10.1016/s0735-1097(20)32192-6).
53. Zeng W, Yuan J, Yuan C, Wang Q, Liu F, Wang Y. Classification of myocardial infarction based on hybrid feature extraction and artificial intelligence tools by adopting tunable-Q wavelet transform (TQWT), variational mode decomposition (VMD) and neural networks. *Artif Intell Med*. Published online 2020. <https://doi.org/10.1016/j.artmed.2020.101848>.
54. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* (80-). Published online 2019. <https://doi.org/10.1126/science.aax2342>.
55. Slater J. Spirituality and the curriculum. *Taboo J Cult Educ*. Published online 2005.
56. New AMA policy recognizes racism as a public health threat. *AMA*. Published 2020. <https://www.ama-assn.org/press-center/press-releases/new-ama-policy-recognizes-racism-public-health-threat>. Accessed 12 Dec 2020.
57. Johnson K. Researchers are starting to refuse to review Google AI papers. *Venture Beat*.
58. Ownby GT. Malpractice case: you're liable, even if your EHR malfunctions. *MedScape*.
59. Makary MA, Daniel M. Medical error—the third leading cause of death in the US. *BMJ*. Published online 2016. <https://doi.org/10.1136/bmj.i2139>.
60. Commission E. A European strategy for data 19.2.2020 COM(2020) 66 Final Communication. 2020.
61. Segal A. The coming tech cold war with china beijing is already countering washington's policy. *Foreign affairs*. Published 9 September 2020. Accessed via <https://www.foreignaffairs.com/articles/north-america/2020-09-09/coming-tech-cold-war-china>. Last access 15 August 2021.
62. Feijóo C, Kwon Y, Bauer JM, et al. Harnessing artificial intelligence (AI) to increase wellbeing for all: the case for a new technology diplomacy. *Telecomm Policy*. Published online 2020. <https://doi.org/10.1016/j.telpol.2020.101988>.



Ayomide Owoyemi, Adenekan Osiyemi, Joshua Owoyemi,
and Andy Boyd

Contents

Introduction	614
Brief History of Artificial Intelligence for Medicine in Africa	615
Current Applications of AI for Medicine in Africa	615
Challenges with AIM Landscape in Africa	616
Digital Health Foundation	616
Data Availability and Quality	616
Infrastructure	617
Costs and Funding	617
Governance, Regulations, and Ethics	617
Critical Areas of Attention for Improving AIM in Africa	617
Governance and Ethical Approaches	618
Building for the Africa (How Best to Approach Solutions and Implementation)	618
Opportunities for AIM Impact in Africa	619
Overview of Selected Applications of AI for Healthcare in Africa	619
Kenya Medical Supplies Agency (KEMSA) with IBM's Watson	620
Afyapap in Zimbabwe	620
Delft Institute's CAD4TB Software	620
References	621

Abstract

The healthcare landscape in Africa has undergone significant changes, thanks to the recent technological advances such as access to the Internet, cloud computing, alternate energy source, and improved smartphone uptake. This adoption of technology as a tool has in recent times included artificial intelligence. The usage of artificial intelligence in medicine

A. Owoyemi (✉) · A. Boyd
University of Illinois at Chicago, Chicago, IL, USA
e-mail: boyda@uic.edu

A. Osiyemi
University of Ibadan Teaching Hospital Ibadan, Ibadan,
Nigeria

J. Owoyemi
Elix Incorporated, Tokyo, Japan

in Africa has seen a few pilots and test cases in human resource planning, child health, diagnostics, and pharmaceuticals. Extant challenges around availability of informative data, legal and policy, costs of development, and lack of infrastructure have stymied progress in different parts of the continent. Despite these challenges, the prospects for AI in Africa to play active and significant roles in population health, individual care, health systems, and pharmaceuticals and medical technology are evident, and solving the highlighted challenges will help make its prospects possible. A major lesson from the experience of AI professionals working in resource-poor settings is that AI implementation should focus on incorporating products into existing systems and institutions rather than attempting to start from scratch or hoping to replace existing systems.

Keywords

Africa · Low-resource · AIM · Primary health · Rwanda · Governance · SDGs · Developing countries

Introduction

Data drives scientific knowledge and advances in healthcare. Different approaches to human research – basic/preclinical research, observational studies, clinical trials, translational studies, and implementation research – have centered on collection, analysis, and interpretation of data that can improve individual and population health. In Africa, productive efforts have been seen with regard to the data revolution. For instance, health research output in Africa has consistently increased in number with some improved quality over the past two decades, especially in areas of infectious diseases and some non-communicable diseases. This increase is in spite of the continued emigration of clinicians and academic professionals away from Africa [1]. The improvement in health research effort is commendable although Africa contributes less than 1% of world research, there is an

opportunity for data analytics can change this low output [2].

The data source and data quality are important factors that determine whether an algorithm can recognize meaningful patterns, learn from experience, and make appropriate decisions given its perceptual and computational limitations. In African countries, most health information utilized for national planning are sourced from household surveys [3]. Some benefits of these surveys, especially those supported by multilateral organizations, include the high level of standardization, good quality control, adept interviewers, and careful data analysis. These surveys have been conducted more frequently in West Africa compared to other parts of the continent, while the central subregion has reported the least surveys over the past 25 years. Other sources of health information in Africa are population census, civil registration systems, disease surveillance, research studies, and health services statistics. These data sources are often fragmented. Generally, data with multiple sources are preferred to those with one source. However, in Africa, mechanisms for triangulation of available data are needed to minimize fragmentation and duplication while ensuring sound data is available in timely ways [4]. More so, computers can simulate more intelligent behaviors and make nuanced decisions with harmonized/pooled data rather than fragmented ones. Data actors including data managers, statisticians, and data analysts together with health departments will need to be involved at various stages of data collection so that developed algorithms can make the best use of data for Africa's population health.

Healthcare landscape in Africa has undergone significant changes, thanks to the recent technological advances such as access to the Internet, cloud computing, alternate energy source, and improved smartphone uptake. These technological innovations have improved the quantity and quality of health data available in the continent as well as the efficiency of data analysis, storage, and utilization as cloud computing continues to gain ground across the continent. In Africa, large volumes of health data are being generated – whether from rural health centers or from some of the

sophisticated genomic laboratories. Advances in information technology and data science have the potential to help resolve some complex developmental issues across the region. This has been recognized by multilateral organizations and research bodies including the World Health Organization (WHO), United Nations International Children's Emergency Fund (UNICEF), and National Institutes of Health (NIH) as well as corporate organizations [5, 6]. The ability to fully extract meaningful knowledge from Africa's health data will lead to rapid innovations that positively impact health of Africans and world at large.

Brief History of Artificial Intelligence for Medicine in Africa

Certain African folklores describe historical stories that bear semblance with modern-day computer science. One that is publicly available is about Ifa, the deity of wisdom, and intellectual development among Yoruba people in Nigeria, West Africa. The Ifa's divination system developed over 12,000 years ago, and still being practiced till date, utilizes binary codes and pattern recognition for communication between an Ifa priest and the gods. This divination system was also applied in solving health problems according to history [7].

Evidence-based application of artificial intelligence in medicine (AIM) was piloted in Africa in the mid-1980s. In 1986, a successful pilot test was initiated in Egypt to improve the detection of common and potentially blinding eye disorders before full-scale deployment in the USA [8]. A probabilistic decision-making system assisted rural health workers to identify life-threatening conditions in Gambian outpatient clinics. The system performed well in detecting 88% of cases among Gambians. In Kenya, AIM improved health worker-patient interaction quality with evidence of increased number of symptoms elicited [9]. Additionally, Computerized Aid To Treat (CATT) was also used in drug prescriptions in South Africa by nurses based on a cost-and-effectiveness algorithm [10].

Current Applications of AI for Medicine in Africa

The most significant applications of AI are happening in business and finance in Africa, but when it comes to AIM, more meaningful applications seem to be occurring in developed nations compared with what is obtained in Africa. Worldwide, it is anticipated that AI can help in saving over \$150 billion in costs by 2026, and different studies have documented the positive impact of AI in delivering benefits for different Sustainable Development Goals (SDGs) including health which is SDG 3 [11]. The prospects of AIM prompted the United Nations in different forums which brought different stakeholders together to discuss how AI can be used to deliver critical public services and help in the journey toward achieving the SDGs [12].

There has been a steady increase in the application of AI as a tool to solve health problems across the continent. Different pilots and test projects are being implemented in countries in the region for health challenges particularly around human resource, health education, diagnosis, consultation, and pharmaceuticals. Examples include a multinomial logistic classifier-based system that predicts the length of stay in public service among health workers and diagnosis of diabetic retinopathy in Zambia, algorithms for authentication of drugs in Nigeria, and chatbot for health education and consultation like the Ada Health and Swahili bot [13, 14].

A significant number of these projects are being led by private organizations and academic institutions which include Delft, IBM, Microsoft, Google, Babylon, etc. While early results from some of these pilots and tests have shown significant promise, there are numerous unanswered questions regarding how much impact these AIM tools have had or are having in terms of healthcare delivery in the continent.

In Africa, radiology seems to be the aspect seeing the most meaningful application of AIM, with the most advanced and integrated uses of AIM happening in that aspect of healthcare delivery. This includes the Delft Institute's software for tuberculosis diagnosis which has been deployed

in Mozambique, Malawi, and Eswatini [15]. Outside diagnostics, health education and information are other areas experiencing significant changes in Africa's AIM. Chatbots are gaining ground in delivering health information, triaging, and consultation to the public. Babylon Health is working with the Rwandan Government to roll out its artificial intelligence-powered triage and symptom checker to the whole country as part of a 10-year partnership [16].

Challenges with AIM Landscape in Africa

While implementation of AIM in Africa will have to face the extant challenges that affect health systems across the continent, there are specific issues that are peculiar to the successful deployment of AIM to meaningfully solve problems. These include economic, social, and institutional problems. These issues concern basic building blocks that are important to help ensure that AIM can be locally designed, developed, tested, and implemented at scale. These will be discussed under the following headings.

Digital Health Foundation

A digitized health system is a necessary foundation for the development and deployment of AIM. This digitalization will include widespread adoption of electronic health record (EHR), efficiency of interoperability, and creation of standards, registries, and infrastructures (e.g., storage).

Digital health has been gaining traction in the region over the past few years with more countries developing and adopting national digital strategies to drive growth in their health sector. However, most of the growth have occurred in the private sector and businesses, while the public sector lags behind due to inefficient governance, poor institutional capacity, and funding [17]. For example, adoption of EHR is limited across the continent where administrative use case is more prevalent than clinical application of EHR. Additionally, most of the clinical uses of EHR occurs in

donor-funded HIV treatment programs. The EHR of these HIV programs are often standalones and are often not integrated into a country's existing health system [18]. This results in difficulties with the creation and curation of health data, while problems with subsequent integration and adoption of AIM systems for health organizations and service providers across the region persist. Building a solid and integrated digital health foundation will pave the way for the smoother development and implementation of AIM.

Data Availability and Quality

Due to the low adoption of EHR systems and inadequate digital infrastructure across the healthcare delivery chain, there is inadequate data to locally train and develop AI systems in the region as most health records are still maintained in paper format [18].

Algorithms being designed for general use globally are trained with data that are not adequately representative of the African populace. This creates inherent bias in these systems that may be deployed to serve the Africans. For example, the International Skin Imaging Collaboration: Melanoma Project, which is one of the largest, open-source archives of pigmented lesions, has most of its data collected from Caucasian populations. Deploying any algorithm trained on this dataset may lead to underperformance among populations with darker skin tones [19]. A study conducted in Uganda using a dermatology AI software showed poor accuracy due to the training data predominantly originating from a Caucasian population [20]. Inadequate meaningful health data has slowed the progress of algorithms that will be more effective among African population. This has led to a reliance on foreign systems for application on the continent. These systems which will be affected by the bias, prejudices, and specific beliefs of the creators lead to unintended bias when applied to low-resource settings [21, 22]. Beyond clinical data, other health data such social determinants of health contribute to a robust AIM system. Some of these nonclinical health information is either still collected

traditionally with pen and paper or exists in inaccessible silos.

Infrastructure

AIM solutions will rely and be built on and integrated with existing digital health infrastructures which are often absent or inadequate in most African countries. African countries lag in Internet penetration and access to electricity (less than 30% of health facilities have electricity access) which has impeded the execution and sustainability of different solutions and technology in healthcare and other sectors [13]. These solutions will have to be deployed on digital devices either for use by health workers or individuals. While uptake of digital devices and adoption of ICT are rising on the continent (smartphone adoption stands at about 40%), most health systems and individuals on the continent still lack access to these devices which will hamper the scaling and adoption of these solutions going forward [23].

Costs and Funding

The healthcare system in the continent suffers from chronic underfunding and massively relies on donor funding. This challenge extends to digital health initiatives and is compounded by the existing unstructured market. Most of the funding for digital health projects including AIM systems come from outside the continent.

It is hard to ascertain how much the development of AIM might cost in developing countries. Considering the approach to development which will involve skills, data acquisition and preparation, and hardware and computing resources alongside system maintenance and upgrading, putting together these systems in Africa will be expensive. African countries are seeing a rise in the acquisition of skills required for development of AI systems as evidenced by the works of Data Science Nigeria and Zindi and the opening of hubs by Google, IBM, and Microsoft but still have a long way to go in getting the required

pool of expertise for local development of these systems [24, 25].

Governance, Regulations, and Ethics

African countries are at different stages in the development, adoption, and implementation of national digital health policies that will guide digital health. These policies and their implementation have been stymied by poor government engagement and political will, thereby negatively impacting the progress of digital health solutions [26].

The legal framework covering data use and protection is just being developed in most African countries. Only nine countries have passed data protection laws across the continent. Considering the repeated history of medical abuse of processes, research, and humans usually by outside actors, these laws are essential to ensuring that the data that will be used to develop AIM systems are ethically obtained and not misused. There is also a need for policies and guidelines that oversee the development and deployment of these AIM systems to ensure that these systems are just, fair, and aligned with Africa's values/heritage with clear lines of accountability. This will minimize the mistrust and hesitation associated with new technology in the region as some locals perceive that external actors are trying to impose their values [27].

Critical Areas of Attention for Improving AIM in Africa

African countries are seeing increasing development of AIM products and services; while the growth is encouraging and promising, the full potential of these will not be realized unless a few things are put in perspective and gotten right. These will be discussed under the following categories:

Data Ecosystem: As AIM will rely on significant volumes of data for training, and it is best that these data are locally generated, it will be important to work on creating systems that enable much easier data generation and curation; this will

involve the digitization of all process along the health delivery chain, widespread adoption of electronic health record, and creation of open data initiatives that increases access to national and government datasets. Rwanda is making progress in this regard as a significant hub for technology in Africa. It signed a deal with Inmarsat to lay the foundation for an Internet of Things (IoT) project to create an infrastructure that allows the easy generation, stability, and sharing of data for the required monitoring and development of solutions. Rwanda also has signed a deal with Babylon Health to create a nationwide digital health delivery system which covers significant digitization of health records. These ensure a foundation to help generate required data and foundation for development of AIM [28, 29].

Infrastructure: Africa has significant infrastructure gaps that cut across all sectors, especially in the aspect of electricity which is most significant to AIM and digital health [30]. While most consumer products might be deployed on smartphones that require less susceptibility to electricity limitations, integration into the health system will depend on access to electricity that will power the digital systems that the AIM will leverage and work on. The stability of electricity will affect the usability of such systems. Alternative power systems like solar electricity are gaining ground and being deployed to ensure power stability for health systems but ensuring good public electricity infrastructure will be more sustainable.

Development and Capacity Building: To enable development of more locally relevant and less bias products, it is important that local actors and stakeholders are fully integrated into the development process and are consulted at every stage. Development of AIM products as the next interesting thing in healthcare will only create tools that will not be useful, efficient, sustainable, and adopted by the stakeholders in the system; this is more important for products that are not directed primarily at consumers.

It is also important to work on capacity building for the respective governments, health workers, and individuals in the respective countries. This will also involve training and development of local talents and skills required to build and maintain these systems. Google has opened an AI

research center in Ghana, while IBM has two research centers, one in Nairobi, Kenya, and another in Johannesburg, South Africa; these research centers are meant to help develop talents in the region and ensure that solutions are developed and tested within the environment where they are needed [31].

Governance and Ethical Approaches

It is important that appropriate governance framework and foundations are laid before AIM gains traction as a regular tool for healthcare delivery across Africa. More countries need to ratify and implement data protection laws to protect against misuse and abuse of data. This also needs to extend to the creation of accountability for the systems that will be created. Carman and Rosman advised that existing principles and frameworks for developing good AI should not be adopted into the African context without holistic evaluation; they further advised that individuals and societies driving research in this area should ensure that values, culture, and contexts are put into consideration while ensuring that knowledgeable stakeholders are involved in all phases of development [27]. Much like everywhere else, it will be helpful for the African Union and her member states to create a committee to design ethical guidelines and frameworks for the implementation of AI in Africa as the European Union did.

Building for the Africa (How Best to Approach Solutions and Implementation)

- Availability of data being one of the major challenges for AIM, a continental network model for interconnecting nations in Africa through its data centers has been proposed. This will involve building world-class infrastructures connecting different medical facilities for collaboration and resource sharing, with the benefit of fast-tracking collaboration and speeding up data collection, storage, and processing in healthcare services [32]. Federated learning, a collaborative machine learning

without centralized training data, could be a viable solution to the problem of fragmented data [33]. Even though this technology was originally proposed mainly to ensure data safety and anonymity, it promises to also act as a method to build robust AI systems over decentralized data sources. This in effect cuts the cost of storing data in the cloud, in a single location. It leverages the diversity in data variation and preserves the integrity of each source, allowing for smarter models, lower latency, and lesser power consumption.

- Internet penetrations is rising in Africa, but a lot of places where these solutions might be relevant either have no access to the Internet or have bandwidth challenges. The solutions (especially ones targeted for use by frontline staff) developed need to make considerations and contingency plans for this challenge. This might involve the development of systems that can run offline and update when it has intermittent access to the Internet or low-end products that do not require large bandwidth to run appropriately.
- There are funding limitations, and it is best that solutions are targeted at relevant and high-impact challenges. These solutions should be funded and developed based on evidence and not as a “shiny new thing” concept. It is also important that health workers, targeted users, and all relevant stakeholders are involved in the design process to ensure that the solution is appropriately targeted at the problem. Most importantly, it is best that the development of these solutions is led by Africans.
- AIM is a tool and not an elixir; therefore, solutions should be built with the state of the health systems and governance in mind and with an understanding that they will be subject to the present challenges that are inherent in these targeted health systems.

Opportunities for AIM Impact in Africa

Considering the dominant health challenges on the continent and the state of the health sectors across different areas of the country, AIM will best be developed and deployed as solutions

under the categories based on a US Agency for International Development (USAID) report on the use of AI in global health which classified AI application in global health under four specific use case areas that are likely to have the highest potential for impact on global health. We will be expanding that to cover five areas by adding another significant use case that is important to healthcare delivery on the continent. These five areas are also relevant to primary healthcare which is very integral to improving health outcomes on the continent. These classes are:

AI-Enabled Population Health – This covers solutions that accept and analyze population health data and provide recommendations based on them. Examples include disease surveillance and warning systems, automated triaging solutions, etc.

Patient Virtual Health Assistance – This covers solution that assists patients on self-care, wellness, and health advice. Examples include health chatbots, screening tools, behavior change tools, etc.

Frontline Health Worker Virtual Health Assistance – This involves solutions that assist frontline healthcare workers’ expertise in co-ordinating patient care, education, advice, and required diagnosis. Examples include screening tools, AI-assisted diagnostics, and health record management.

Physician Clinical Decision Support – This covers the provision of specialist expertise to primary care physicians and generalists who are most involved in clinical care across the continent. Examples include AI-assisted diagnostics, clinical decision support, and quality assurance and training.

AI-Assisted Logistics – This class covers solutions that help to address and manage the supply chain of drugs and other essential health products [34].

Overview of Selected Applications of AI for Healthcare in Africa

Below are three selected cases of the application of AIM in Africa that are worthy of note. They

cover diagnosis, supply chain management, and health education, three of the most significant areas where AI will be strategic in making a difference in healthcare on the continent.

Kenya Medical Supplies Agency (KEMSA) with IBM's Watson

The Kenya Medical Supplies Authority (KEMSA) is a state-owned health logistics service company founded in the year 2000. This company has the authorization to procure, store or warehouse, and distribute medical commodities to public healthcare facilities and other public sector customers in Kenya. To improve on its service delivery, KEMSA partnered with IBM's Watson.

IBM's Watson is used by the Kenya Medical Supplies Authority (KEMSA) at its approximately 7000 facilities across the country. Pharmacists, patients, and doctors can interact with the AI through SMS, voice messages, and online messaging services. It can instantaneously access information and is being used to aid the medical supply chain. It can also act as an advisor that prevents supply shortages and finds new suppliers if needed. The health workers can find out if there is a medical supply available at their facility, when the next shipment will arrive, and if their hazardous wastes are disposed of. It is also used to maximize patient benefits as it allows them to find out which clinics have their required medications and how to properly dispose medical wastes at home. Users can interact with the AI through various platforms, including SMS, computer, and voice over mobile data [35].

Afya Pap in Zimbabwe

Afya Pap is a mobile app that serves as a personalized health education and coaching tool on chronic diseases especially diabetes and high blood pressure. It uses artificial intelligence and behavioral science to deliver personalized health advice and health insights to users. Afya Pap takes the medical conditions of the users and their lifestyle and culture into consideration, and based on these, the app can provide daily customized health

tips varying from easy exercise tips to medication reminders or eating healthier local foods. Also, users have an avatar that is a visual representation of their health status. The more they follow the provided health advice, the healthier their avatar looks.

Afya Pap is available for use in seven countries in Africa including Zimbabwe, Kenya, Uganda, Egypt, and others. It is a subscription-based app as users must pay a monthly subscription fee to use after the free 14-day trial period expires. Although Afya Pap was created to help patients suffering from diabetes and high blood pressure, it was extended at the beginning of the COVID-19 pandemic to provide live consultations with local physicians while also delivering COVID-19 prevention and care messaging to curb the spread of misinformation [36].

Delft Institute's CAD4TB Software

CAD4TB (Computer-Aided Detection for Tuberculosis) was designed to reduce costs and time involved in diagnosing tuberculosis. It was developed by Delft's Institute in the Netherlands; CAD4TB interprets digital x-rays using deep learning methods and remote expertise. It solves two critical problems in the diagnosis of tuberculosis.

- It delivers its results within a record time of 60 s to 1 min allowing patients to be diagnosed and treated within a day. Statistics have shown that early diagnosis and treatment of tuberculosis can not only save millions of lives but also curb its spread and reduce death rate from TB.
- The software beats other tests in cost-effectiveness. Aside from the fact that it is relatively cheap, it shows who is more likely to have tuberculosis and will need further testing, thus reducing the number of people that must spend a lot of money on the more expensive tests.

The CAD4TB can be used with or without an Internet connection. When a good Internet connection is available, the x-ray is processed by the CAD4TB server in the cloud. This allows it to be shared with other professional healthcare

providers who can provide expert advice remotely. In the absence of an Internet connection, the CAD4TB box which is connected to the x-ray system analyzes the image [37, 38].

There is a significant potential offered by AIM in improving healthcare in Africa. The existing use cases show that it is a feasible tool for tackling health challenges, reducing costs, and improving health access and quality. An evidence-based approach should be adopted in decision-making and implementation of AIM in Africa. African countries and institutions must also enact laws, guidelines, and policies that will ensure that users are protected and the products are developed for impact. A major lesson from the experience of AI professionals working in resource-poor settings is that AI implementation should focus on incorporating products into existing systems and institutions rather than attempting to start from scratch or hoping to replace existing systems.

References

1. Pang T, Lansang MA, Haines A. Brain drain and health professionals. *BMJ*. 2002;324(7336):499–500.
2. Elsevier. Africa generates less than 1% of the world's research; data analytics can change that [Internet]. Elsevier Connect. [cited 2021 Feb 17]. <https://www.elsevier.com/connect/africa-generates-less-than-1-of-the-worlds-research-data-analytics-can-change-that>
3. Mbondji PE, Kebede D, Soumbey-Alley EW, Zielinski C, Kouvidilwa W, Lusamba-Dikassa P-S. Health information systems in Africa: descriptive analysis of data sources, information products and health statistics. *J R Soc Med*. 2014;107(1 Suppl):34–45.
4. Kadengye DT. Why fixing Africa's data gaps will lead to better health policies [Internet]. The Conversation. [cited 2021 Feb 17]. <http://theconversation.com/why-fixing-africas-data-gaps-will-lead-to-better-health-polices-111869>
5. Harnessing Data Science for Health Discovery and Innovation in Africa [Internet]. 2019 [cited 2021 Feb 17]. <https://commonfund.nih.gov/africadata/funding>
6. Hoodbhoy Z, Hasan B, Siddiqui K. Does artificial intelligence have any role in healthcare in low resource settings? *J Med Artif Intell* [Internet]. 2019 Feb 7 [cited 2019 Nov 24];2(0). <http://jmai.amegroups.com/article/view/5049>
7. Alamu F, Aworinde H, Isharufe W. A comparative study of Ifa divination and computer science. *Int J Innov Technol Res*. 2013;1:524–8.
8. Kastner JK, Dawson CR, Weiss SM, Kern KB, Kulikowski CA. An expert consultation system for frontline health workers in primary eye care. *J Med Syst*. 1984;8(5):389–97.
9. Hunter J, Cookson J, Wyatt J, editors. *AIME 89: second European conference on artificial intelligence in medicine*, London, August 29th–31st 1989. Proceedings [Internet]. Berlin/Heidelberg: Springer Berlin Heidelberg; 1989 [cited 2020 Apr 25]. (Rienhoff O, Lindberg DAB, editors. Lecture Notes in Medical Informatics; vol. 38). <http://link.springer.com/10.1007/978-3-642-93437-7>
10. Forster D, International Development Research Centre (Canada). Expert systems in health for developing countries: practice, problems, and potential. Ottawa: International Development Research Centre; 1992.
11. Forbes Insights: AI And Healthcare: A Giant Opportunity [Internet]. [cited 2019 Nov 24]. <https://www.forbes.com/sites/insights-intelai/2019/02/11/ai-and-healthcare-a-giant-opportunity/#3afb11224c68>
12. United Nations Activities on Artificial Intelligence (AI). 66.
13. Owoyemi A, Owoyemi J, Osiyemi A, Boyd A. Artificial intelligence for healthcare in Africa. *Front Digit Health* [Internet]. 2020 [cited 2020 Jul 8];2. <https://www.frontiersin.org/articles/10.3389/fdgth.2020.00006/full>
14. Newey S. World's first AI health app in Swahili launches to tackle doctor shortages. *The Telegraph* [Internet]. 2019 Nov 19 [cited 2021 Feb 17]; <https://www.telegraph.co.uk/global-health/science-and-disease/worlds-first-ai-health-app-swahili-launches-tackle-doctor-shortages/>
15. Our Projects [Internet]. Delft Imaging. [cited 2021 Feb 17]. <https://www.delft.care/projects/>
16. Rwanda will be the world's most advanced country for digital health [Internet]. Babylon Health. [cited 2021 Feb 17]. <https://www.babylonhealth.com/blog/business/rwanda-will-be-the-worlds-most-advanced-country-for-digital-health>
17. Trends in digital health in Africa: Lessons from the African strategies for health project [Internet]. U.S. Agency for International Development (USAID); 2016 Sep [cited 2021 Feb 17]. https://www.msh.org/sites/default/files/digital_health_in_depth_review_final.pdf
18. Odekunle FF, Odekunle RO, Shankar S. Why sub-Saharan Africa lags in electronic health record adoption and possible strategies to increase its adoption in this region. *Int J Health Sci (Qassim)*. 2017;11(4): 59–64.
19. ISDIS [Internet]. [cited 2021 Feb 17]. <https://isdis.org/>
20. Kamulegeya LH, Okello M, Bwanika JM, Musinguzi D, Lubega W, Rusoke D, et al. Using artificial intelligence on dermatology conditions in Uganda: a case for diversity in training data sets for machine learning. *bioRxiv*. 2019;826057.
21. Wahl B, Cossy-Gantner A, Germann S, Schwalbe NR. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? *BMJ Glob Health*. 2018;3(4):e000798.
22. Mahomed S. Healthcare, artificial intelligence and the fourth industrial revolution: ethical, social and legal

- considerations. *S Afr J Bioethics Law*. 2018;11(2): 93–5.
23. Africa's Smartphone Market Posts Growth, but Uncertainty Around Global COVID-19 Outbreak Casts Shadow over Short-Term Prospects [Internet]. IDC: The premier global market intelligence company. [cited 2021 Feb 17]. <https://www.idc.com/getdoc.jsp?containerId=prMETA46110420>
24. Zindi rallies Africa's data scientists to crowd-solve local problems [Internet]. TechCrunch. [cited 2019 Nov 24]. <http://social.techcrunch.com/2019/08/09/zindi-rallies-africas-data-scientists-to-crowd-solve-local-problems/>
25. My Life In Tech: Olubayo Adekanmbi is using his privilege to change the world [Internet]. TechCabal. 2019 [cited 2019 Nov 24]. <https://techcabal.com/2019/10/23/my-life-in-tech-olubayo-adekanmbi-is-using-his-privilege-to-change-the-world/>
26. ARTIFICIAL INTELLIGENCE: Starting the policy dialogue in Africa [Internet]. World Wide Web Foundation; 2017 Dec [cited 2021 Feb 17]. <http://webfoundation.org/docs/2017/12/Artificial-Intelligence-starting-the-policy-dialogue-in-Africa.pdf>
27. Carman M, Rosman B. Applying a principle of explicability to AI research in Africa: should we do it? *Ethics Inf Technol* [Internet]. 2020 May 11 [cited 2021 Feb 17]; <https://doi.org/10.1007/s10676-020-09534-2>
28. Jack A. Rwanda venture tests digital health potential in developing world [Internet]. 2021 [cited 2021 Feb 17]. <https://www.ft.com/content/4fe33c92-cbd5-459a-8df6-20d0d1f57ec8>
29. Case study: Smart City Kigali, Rwanda [Internet]. Inmarsat Corporate Website. [cited 2021 Feb 17]. <https://www.inmarsat.com/en/insights/enterprise/2017/case-study-smart-city-kigali-rwanda.html>
30. Madden GJ and P. Figures of the week: Africa's infrastructure needs are an investment opportunity [Internet]. Brookings. 2019 [cited 2021 Feb 17]. <https://www.brookings.edu/blog/africa-in-focus/2019/06/27/figures-of-the-week-africas-infrastructure-needs-are-an-investment-opportunity/>
31. The future of AI research is in Africa [Internet]. MIT Technology Review. [cited 2021 Feb 17]. <https://www.technologyreview.com/2019/06/21/134820/ai-africa-machine-learning-ibm-google/>
32. Ajayi OO, Bagula AB, Maluleke HM. Africa 3: a continental network model to enable the African Fourth Industrial Revolution. arXiv:201012020 [cs] [Internet]. 2020 Oct 14 [cited 2021 Feb 17]; <http://arxiv.org/abs/2010.12020>
33. Federated Learning: Collaborative Machine Learning without Centralized Training Data [Internet]. Google AI Blog. [cited 2021 Feb 17]. <http://ai.googleblog.com/2017/04/federated-learning-collaborative.html>
34. Artificial Intelligence in Global Health: Defining a Collective Path Forward [Internet]. Washington, DC: USAID, The Rockefeller Foundation, Bill & Melinda Gates Foundation; 2019 [cited 2019 Nov 24]. https://www.usaid.gov/sites/default/files/documents/1864/AI-in-Global-Health_webFinal_508.pdf
35. Solving Africa's healthcare logistics problems with AI [Internet]. [cited 2021 Feb 21]. <https://www.bizcommunity.com/Article/196/159/179698.html>
36. AI and Behavioral Science to Manage Diabetes in Africa [Internet]. Global Innovation Path. 2018 [cited 2021 Feb 21]. <https://globalinnovationpath.com/bao-bab-circle/>
37. X-ray Systems [Internet]. Delft Imaging. [cited 2021 Feb 21]. <https://www.delft.care/x-ray-systems/>
38. Khan FA, Pande T, Tessema B, Song R, Benedetti A, Pai M, et al. Computer-aided reading of tuberculosis chest radiography: moving the research agenda forward to inform policy. *Eur Respir J* [Internet]. 2017 Jul 1 [cited 2021 Feb 21];50(1). <http://erj.ersjournals.com/content/50/1/1700953>



Aim in Climate Change and City Pollution

45

Pablo Torres, Beril Sirmacek, Sergio Hoyas, and Ricardo Vinuesa

Contents

Introduction	624
Machine-Learning Methods in the Study of Urban Pollution	625
ML Methods in Air-Pollutant Modeling	625
ML Methods to Model Flow Dynamics	628
Remote Sensing for Urban Air Observation	629
Impact of Remote-Sensing Sensors for Monitoring Urban Airflow	629
Remote-Sensing Data Resources and Analysis Methods	630
Further Supportive Data that Satellite Remote Sensing Can Offer	631
Challenges and Open Problems	632
References	633

Abstract

The sustainability of urban environments is an increasingly relevant problem. Air pollution plays a key role in the degradation of the environment as well as the health of the citizens exposed to it. In this chapter we provide a review of the methods available to model air pollution, focusing on the application of machine-learning methods. In fact, machine-learning methods have proved to importantly increase the accuracy of traditional air-pollution approaches while limiting the development cost of the models. Machine-learning tools have opened new approaches to study air pollution, such as flow-dynamics modeling or remote-sensing methodologies.

Supplementary Information: The online version of this chapter (https://doi.org/10.1007/978-3-030-64573-1_290) contains supplementary material, which is available to authorized users.

P. Torres · S. Hoyas
Instituto Universitario de Matemática Pura y Aplicada,
Universitat Politècnica de Valencia, Valencia, Spain
e-mail: pablotg@kth.se

B. Sirmacek
Smart Cities, School of Creative Technologies, Saxion
University of Applied Sciences, Enschede, The
Netherlands
e-mail: b.sirmacek@saxion.nl

R. Vinuesa (✉)
FLOW, Engineering Mechanics, KTH Royal Institute of
Technology, Stockholm, Sweden
e-mail: rvinuesa@mech.kth.se

Keywords

Remote sensing · Satellite image processing · Earth observation · Air quality · Urban airflow · Machine-learning · Urban flows · Fluid mechanics · Climate change

Introduction

Urban areas are at the center of the current climate-change debate. Starting at the end of the industrial revolution, urban areas have been growing at an accelerated rate. By 2050 the European Commission expects 70% of the global population to live in urban environments [4]. Thus, it is clear that cities play and will continue to play a major role in our society. Cities are responsible for a great share in the acceleration of climate change.

Mainly due to anthropogenic activity, cities act both as concentrators and diffusers of contaminants at the expense of not only the city itself but also of neighboring areas. In fact, ambient air pollution is responsible for 790,000 deaths in Europe [11], as shown in more detail in Fig. 1. Furthermore, the vast majority of cities tend to act as the so-called urban-heat-islands (UHI), modifying the natural thermal dynamics of neighboring areas. For those reasons among others, cities are expected to play a key role in the policies and endeavors aiming to reverse the effects of climate change within the next decade [18].

To deal with pollutant dispersion within the urban environment, one inevitably needs to consider the dynamics of urban flows. Environmental sciences tend to rely on the use of models to study and predict the behavior of atmospheric flows. Models can provide very useful

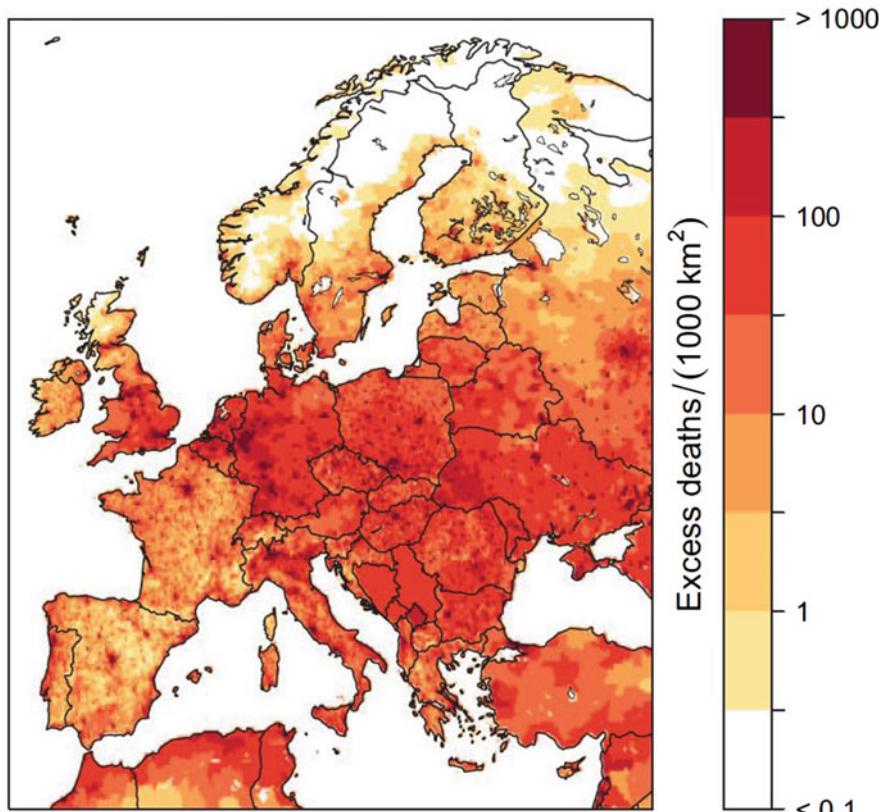


Fig. 1 Distribution of annual excess mortality from cardiovascular diseases due to air pollution in Europe. (Figure reproduced from Lelieveld et al., with permission from the publisher (Oxford University Press))

information when studying atmospheric phenomena at a very general level or when dealing with specific applications. However, to understand the flow within an urban environment as well as the inner relation to pollution mechanisms, modeling appears to be limited. In fact, urban flows are characterized by their complexity, thus severely difficulting the use of models to study the effects on pollutant dispersion or thermal problems. The experimental approach, although it can bring very useful insight into specific applications, typically does not provide information on the physical mechanisms driving urban flows [17]. Numerical simulations are also typically used in the study of turbulent flows and thus can be applied within the context of urban environments. Numerical approaches vary in terms of accuracy and complexity. The so-called Reynolds-averaged Navier-Stokes simulation (RANS) is widely used for its simplicity and reduced computational cost. Unfortunately, RANS methods exhibit important limitations when dealing with complex turbulent flows such as urban flows, thus limiting its use within the context of urban environments. In this way, urban flow simulations have to rely on more accurate approaches, such as direct numerical simulations (DNS) and large-eddy simulations, which have high computational cost. In addition, the domains considered in urban environments are typically large and complex, which again increase the computational cost.

The use of artificial intelligence (AI), and in particular of machine-learning methods, can improve the aforementioned limitations of the computational approach in the study of urban flows. In fact, using machine learning techniques one can produce turbulent models that are significantly more precise than the models historically used in computational fluid dynamics (CFD) without reaching the computational cost of DNS and LES methods. Those models would increase the forecasting capabilities of numerical methods in the study of urban flows. The next section will focus on explaining the state of the art of the aforementioned methods as well as the lines of study currently being explored.

Machine-Learning Methods in the Study of Urban Pollution

Machine-learning (ML) techniques have been widely used in many fields of physics and scientific research. In fact, ML has proved to be very effective in extracting underlying patterns and correlations within complex physical processes. The present section will focus on introducing the different ML approaches in the study of urban pollution. When dealing with urban environment pollution, two approaches are found. On the one hand, some studies focus on modeling pollutant dispersion as an independent physical phenomenon. Those models have been widely used and typically rely on empirical relations. Using ML methods, the aforementioned models can be significantly improved, as we will see later on. On the other hand, many studies rely on flow dynamics to characterize pollutant dispersion. This approach typically consists of solving the behavior of the flow in an urban environment in order to model the behavior of the pollutant. Several numerical methods have been widely used to solve urban flow dynamics (e.g., RANS and LES) but they are either too simplistic or too expensive in terms of computational cost. In this particular area, ML methods are being developed in order to reproduce flow dynamics without having to incur such computational costs.

ML Methods in Air-Pollutant Modeling

Air-pollution models have been widely used in order to analyze and control air quality in cities. Air quality is typically measured in terms of the particulate matter (PM) concentration, which basically measures the amount of particles found in the air. PMs are classified in terms of the size of the particulate. For instance, PM 10 refers to particles with diameters of 10 micrometers or smaller. In the present section we will focus on PM 2.5 as it is common to urban environments and it has been associated both with adverse health effects [21] and climate forcing [1]. Pollutant models have traditionally relied on linear regression techniques to capture the underlying mechanisms of

pollution. This approach is known to be limited, since urban-flow phenomena are characterized by their complexity, that is, the presence of nonlinearities and highly correlated effects. Complex phenomena are known to be challenging to linear methods as they tend to misrepresent important underlying relations [20]. ML arises as a solution to this problem and thus enhances air-pollution models. The fundamental idea behind ML pollutant models lies on using data, typically obtained with on-site measurements, to train the model such that it can predict a certain quantity. In the case of pollution models, the output of the model is normally pollutant concentration, that is, PM concentration. In the next lines we will present the most common ML algorithms applied to urban pollution models and compare them in terms of performance and accessibility.

The random-forest (RF) algorithm is a popular ML method to solve classification and regression problems. It is based on a classical decision-tree algorithm widely used in statistics and ML, but it builds each individual tree from random subsamples of the original data. In this way, each sub-sample of the original data produces a tree and thus produces a predicted response (or a predicted class if dealing with a classification problem). The final output of the model is determined by the average of each of the predictive responses (or classes) determined by each of the data subsamples. The use of a diverse database increases the probability of success [6], thus being very convenient to analyze complex phenomena which typically have important disparities within batches of data. In addition, thanks to the averaging process the algorithm can handle missing values in the dataset [3]. The accuracy of the method is dependent on the strength of each individual tree and the dependency of each subsample with one another [6].

Boosted regression algorithms are typically the result of modifying decision trees regression algorithms in order to enhance performance. For instance, the boosted regression tree (BRT) combines the well-known classification and regression trees (CART) with a large number of single models in order to improve the predictive performance of the model [6]. In fact, the averaging

process found in the RF model is replaced with a forward stagewise procedure in which existing trees are left untouched with new tree development using the residual information generated in the previous step of the process [6]. Alternatively, one can find gradient boosting regression methods such as XG Boost, which use gradient boosted decision trees in classification and regression problems, thus enhancing the speed and performance of the algorithm. In addition, XG Boost can be run using parallel trees as well as a cluster of computers during training time [3], which optimizes the method.

Multilayer perception (MLP) regression is part of the artificial neural network (ANN) family of methods. ANN methods are designed to mimic the behavior of the human brain during the learning process, that is, using interconnected synaptic neurons capable of learning and storing information about their environment [14]. The fundamental component of ANNs are neurons, which are described by a linear combination of weighted input signals and an activation function that limits the amplitude range of the output. ANN is built using at least three layers of neurons: the input, hidden, and output layers. The method processes information sequentially throughout the layers starting at the input layer and finishing at the output layer. The input of the system consists of the training data and a goal to fulfill. In this way, the system is trained on historical data to fulfill the predefined objective, learning during the process the underlying relations that lead to the fulfillment of the goal. Note that several training algorithms can be applied to a given network, for example, the back-propagation algorithm which combine forward and backward passes to train the network [14]. However, taking into account the motivation of the present chapter, we will neglect the training algorithm selection and focus exclusively on the performance of the models once trained. Other, more sophisticated types of ANN include recurrent neural networks (RNNs) and convolutional neural networks (CNNs), which have been used in the context of temporal predictions [13] and non-intrusive sensing [7] of turbulent flows. These strategies have high potential when it comes to developing robust air-pollution frameworks.

The aforementioned techniques are widely used in classification and regression problems and thus their application to air-pollution models – which are regression problems – is currently being studied. The present paragraph aims at discussing the performance of the different ML methods introduced early on. To do so, we will compare each model's performance with one another as well as with linear regression methods used in air-pollution models. To measure performance, four metrics are typically used in the related literature: the mean absolute error (MAE), the root mean square error (RMSE), and the mean square error (MSE). The MAE is defined as the sum of the absolute differences between the predicted values and the actual values scaled with the inverse of the number of observations. Similarly, the MSE is defined as the sum of the squares of the differences between the predicted and actual values scaled with the inverse of the number of observations. The RMSE is just the square root of the MSE. Wang et al. [20] used the aforementioned quantities to compare the performance of two ML methods, the ANN and the XG Boost, with the land use regression model (LUR). The authors developed a least-square LUR by fitting an intercept-only model and then adding explanatory variables (one at a time) based on the ranking of their correlation with the log-transformed of the PM 2.5. Some of the explanatory variables include: land uses, length of neighboring roads, distance to major aerial routes, traffic information, meteorology, etc. The models were developed using air-quality data (PM 2.5) collected over the course of 4 weeks between March and June 2019 in downtown Toronto (Canada). The samples were obtained on a non-rainy weekday between 7:00 AM and noon, covering 19 unique corridors in a 4-by-6 km area [20]. The authors explain that both LUR and ML methods present strengths and flaws. For instance, they found that LUR methods were highly dependent on the selection of the explanatory variables, which are directly dependent on the user's subjective judgment and a priori knowledge. In this way, a change in the selected explanatory variables, *ceteris paribus*, would have an important effect on the performance metrics (MAE, RME, RMSE, etc.). However, when

properly fit, the performance of LUR methods significantly improves – even surpassing ML methods – especially when the size of the dataset decreases [20]. In fact, the size of the dataset is a critical limitation of ML methods since large amounts of data need to be processed within the network's layers such that the model can produce accurate results. Nevertheless, the ANN superiority with respect to linear regression models is guaranteed by the universal approximation theorem of functions, which states that a fully-connected multilayer feedforward neural network with continuous, bounded and nonconstant activation function can act as a universal approximator for any smooth mapping to any accuracy [20]. The superiority of ANNs over traditional linear methods in the context of turbulent flows has also been thoroughly discussed by Guastoni et al. [7]. The authors conclude that ML methods provide opportunities to understand complex underlying relations in air-pollution dispersion processes as well as nonlinear relations between air quality and exogenous variables. Nevertheless, one cannot directly discard regression methods, since they may exhibit very good performance in the cases where local knowledge is available.

Doreswamy and Yogesh [3] presented a similar study where four ML models (RF, BRT, MLP, and CART) were compared in terms of performance. In addition, they also compared ML methods with classical statistical methods such as the linear regression. Once again, performance is analyzed using the error metrics (MAE, RME, RMSE, etc.) compared both in training and testing results. The dataset used in the process was taking hourly measurements of air pollution (PM_{2.5}) at 76 stations distributed over Newport (Taiwan) between 2012 and 2017. The obtained dataset was used to train the different ML algorithms and to compare the predicted air-pollution values with the actual measurements taken in the stations. The gradient-boosting regression – a type of boosting-regression algorithm – showed the best performance compared with both classical statistical methods and ML algorithms. In training, the gradient boosting regressor showed an MSE at one order of magnitude lower than the rest of

ML methods and more than two orders of magnitudes lower than the linear regression method. In the rest of the metrics, the obtained values were two to three times lower than the rest of methods. In tests, the difference between the methods was not as large, but the gradient-boosting regressor still obtained errors two times lower than the rest of the methods. In general, all the analyzed ML algorithms exhibited errors two times lower than the ones obtained with the linear regression [3], thus supporting the superiority of ML methods.

In conclusion, we have discussed different ML methods that can be used to develop air-pollution models. For the majority of the algorithms the performance was found to be better than that of classical statistical methods, such as the linear regression. In this way, the application of ML techniques to develop air-pollution models appears to be useful as the accuracy of the models is significantly improved. However, it is important to keep in mind that ML algorithms are limited by the amount of data needed during training, thus not being applicable when large batches of data are not available.

ML Methods to Model Flow Dynamics

The physics of the flow determines the dynamics of pollutant dispersion. In the previous paragraph we approached the study of air quality by modeling pollution, that is, using particulate matter. An alternative approach consists of studying the dynamics of the flow in order to determine the dispersion of pollutants. Numerical simulation such as LES and DNS can be very useful in order to study the dynamics of the flow. However, this kind of numerical technique has a very important computational cost which limits its use. In addition, the setup of the simulation is not a simple task, which again increases the development time. ML methods could significantly improve the current state of numerical simulations in urban flows by improving computational cost and by extension computational time. The main idea is to use the data obtained from a numerical simulation to predict the future behavior of the flow. Note that this approach is relatively new and

thus the available literature on the matter is limited. Nevertheless, the aforementioned approach is promising, as shown by Srinivasan et al. [13] in the context of recurrent neural networks. Non-intrusive sensing is another area with very high potential for ML-based methods, as shown by Guastoni et al. [8].

Xiao et al. [22] developed a fast-running non-intrusive reduced-order model (NIROM) to predict the behavior of the flow in urban environments. The authors gathered the data obtained from a high-fidelity numerical simulation (LES) in order to create the starting dataset. They used the data to generate snapshots of the flow fields such that it can be processed later on. A singular value decomposition (SVD) was applied simultaneously to all velocity components such that the natural correlation between the components is captured. The result of the process is a series of proper orthogonal decomposition functions, which are then used in a Gaussian process regression (GPR). The GPR consists of applying a linear combination of Gaussian-shape basis functions in order to obtain surface functions, which are one of the necessary inputs to build the network. Using the aforementioned elements, a neural network was trained in order to predict the behavior of the flow governing equations. The main idea is that one uses a dataset obtained using a high-fidelity simulation – which is computationally very costly – to obtain a model (NIROM) that will be able to predict the behavior of the flow, this time without the need of the dataset. The authors used the NIROM to predict the flow around the London South Bank University such that the prediction could be compared with actual measurements of the flow. They found that the NIROM was capable of making predictions beyond the range of the snapshots thus being able to accurately represent the vast majority of the dynamics represented in the high-fidelity model [22].

The main advantage of the NIROM over other ML-methods lies on the small amount of data required to make it function. Recall the gradient-boosting method presented in the previous paragraph, it required a huge amount of data to produce a reliable output. On the contrary, NIROMs are able to produce an accurate result with a

smaller database which can be obtained using a numerical simulation. Comparing the NIROM with high-fidelity numerical simulations, the main advantage of the NIROM lies on computational cost. Once the network is trained, the cost of producing new results is very small compared to the cost of running an additional high-fidelity numerical simulation such as a LES or a DNS. Nevertheless, the NIROM here presented, solves the behavior of the flow and thus needs to be adapted to study pollutant dispersion. This can be done either by including an additional equation in the NIROM (passive scalar) or by running an additional model that uses as input the solution of the governing equation of the flow.

Remote Sensing for Urban Air Observation

Impact of Remote-Sensing Sensors for Monitoring Urban Airflow

The value of remote-sensing data has been noticed in the early urban airflow and urban air-quality studies. Researchers have found opportunities to measure, visualize, and explain urban airflow and quality using remote-sensing images acquired from satellites. Early studies, however, have used the satellite images mostly as a background map in order to visualize the data which is collected from ground sensors. This is mainly because it took relatively longer time to develop higher resolution satellite sensors and to design new algorithms which can extract information about the urban airflow and air quality.

The newer satellite sensors have provided opportunities to researchers to develop algorithms to predict the airflow speed and direction, land surface temperature (LST), and atmospheric aerosol particles or in other words particulate matter (PM) which indicate the micrometer size of the particles that pollute air.

Accurate and real-time measurements of the airflow from satellite sensors still have limitations due to the available sensor specifications, visit frequencies (number of days needed for the satellite trip in the orbit before it returns back to the

same observation point again), spatial resolution, and accuracy of the information. On the other hand, measurement of LST and PM can be done more accurately and frequently with the existing satellite sensors. Therefore, in order to talk about the urban airflow and air qualities frequently LST measurements, PM measurements and the Urban-Heat-Island (UHI) effect are discussed. The UHI effect is the condition that describes higher temperatures in urban areas than surrounding areas of less development. An UHI occurs because of the extensive modification of the land surface. Remarkably higher temperatures are seen mainly in urban areas than in suburban and rural areas, because of the high number of the dense obstacles in urban areas.

Even though the existing remote-sensing satellite sensors cannot be used as high-precision urban airflow measurement tools, they can still provide a number of advantages. The most significant of them could be listed as follows:

- The ground sensors can be used for high-resolution airflow and air-quality measurements. However, the uneven distribution and limited installation possibilities create challenges to collect data which can fully represent the air measurements of the all urban area. On the other hand, using the ground measurements as control points, satellite measurements can be used to estimate the LST and PM for the areas where in situ measurements are not available.
- Satellite images can be used to find the optimal sensor positions in order to install the in situ sensors in urban areas.
- Remote-sensing images can help with monitoring very large areas and make it possible to do large-scale urban airflow research possible.
- Most of the time, it is possible to find digital maps of cities which bring information about the street tunnels, building boundaries, vegetation areas. However, satellite images can provide further information (low/high vegetation type, temperature changes of the lakes and other water reserves, building rooftop types, traffic density, etc.). These details which are acquired from the satellite images give chances to highly enhance the

information for doing detailed urban airflow and air quality analysis.

Remote-Sensing Data Resources and Analysis Methods

Satellite sensors can be classified as active and passive sensors. Active sensors (i.e., radar, LIDAR, etc.) first send a wave and generate an image by using the measurements of the waves which are scattered back. Passive sensors (i.e., multispectral, hyperspectral, thermal, etc.), however, generate images by using the light which is naturally received by the sensor. An active sensor, radar, can be used to measure wind vectors over the ocean through radar backscatter, which is also called scatterometry technique. The scatterometer on the SEASAT satellite makes it possible to resolve the wind direction within a 180-degree directional ambiguity, which is then resolved by knowledge of the overall atmospheric pressure pattern. The SEASAT scatterometer was an outstanding success, pointing the way toward future measurements of ocean wind speed and direction. Unfortunately this satellite had a massive power failure 6 months after it started operation. A replacement, stand-alone satellite, QuikSCAT, was launched quickly thereafter in 1999. The scatterometer on QuikSCAT is known as SeaWinds.

Passive sensors cannot measure the wind vectors straightforward like active sensors can measure. However, indirectly they allow researchers to estimate the airflow by measuring LST and PM. Therefore, they are frequently used in urban air flow studies. One of the most frequently used satellite sensors is the Terra satellite which was launched in December 1999. Terra is a highly valuable satellite which has been serving over 20 years for observing our earth. Terra satellite, which is approximately the size of a small school bus, carries five different sensors to observe different qualities of our earth. These sensors are: Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER), Clouds and Earth's Radiant Energy System (CERES), Multi-angle Imaging Spectroradiometer (MISR),

Measurements of Pollution in the Troposphere (MOPITT), and Moderate Resolution Imaging Spectroradiometer (MODIS). ASTER is the only high spatial resolution instrument on the Terra platform. The Advanced Spaceborne Thermal Emission and Reflection Radiometer obtains high-resolution (15 to 90 square meters per pixel) images of the Earth in 14 different wavelengths of the electromagnetic spectrum, ranging from visible to thermal infrared light. Scientists use ASTER data to create detailed maps of land surface temperature, emissivity, reflectance, and elevation. ASTER provides high-resolution images in 14 different bands of the electromagnetic spectrum, ranging from visible to thermal infrared light. The resolution of images ranges between 15 and 90 meters. ASTER data are used to create detailed maps of surface temperature of land, emissivity, reflectance, and elevation [5]. Because the surface UHIs are typically characterized by (LST), it makes sense to measure LST from remotely sensed data to study UHI effects.

With its sweeping 2330-km-wide viewing swath, MODIS however sees every point on our world every 1–2 days in 36 discrete spectral bands. Consequently, MODIS tracks a wider array of the earth's vital signs than any other Terra sensor. For instance, the sensor measures the percent of the planet's surface that is covered by clouds almost every day. This wide spatial coverage enables MODIS, together with MISR and CERES, to help scientists determine the impact of clouds and aerosols on the Earth's energy budget. MODIS provides daily global coverage and the 10 km resolution of aerosol optical depth (AOD) for studying spatial variability of aerosols in urban areas. AOD is a reasonably good proxy for PM2.5 ground concentrations. Wang et al. [19] used this data to generate air quality maps and compared the information to the acute health needs of the residence in the test area. Téllez-Rojo et al. [16] used the same air-quality mapping approach for Mexico City and showed the direct relation of low air qualities to the acute respiratory symptoms of children.

The MOPITT sensor was designed to enhance our knowledge of the lower atmosphere and to

observe how it interacts with the land and ocean biospheres. The sensor measures emitted and reflected radiance from the Earth in three spectral bands. As this light enters the sensor, it passes along two different paths through onboard containers of carbon monoxide. The different paths absorb different amounts of energy, leading to small differences in the resulting signals that correlate with the presence of these gases in the atmosphere. MOPITT's spatial resolution is 22 km at nadir and it "sees" the Earth in swaths that are 640 km wide. Moreover, it can measure the concentrations of carbon monoxide in 5 km layers down a vertical column of atmosphere, to help scientists track the gas back to its sources.

Another highly used data source comes from the Landsat satellite (passive) sensors. Launched in 1982 on Landsat 4 and 1984 on Landsat 5, researchers found an opportunity to access the TM thermal bands. Landsat 4 operated successfully for over 10 years, with data collection terminated in 1993. The Landsat 5 TM acquired data for over 27 years until communication system failures essentially ended the TM data collections in November 2011. Landsat 6 never reached its operational orbit after launching in 1993. In 1999, Landsat 7 was launched with the ETM+ instrument. The newest Landsat mission, Landsat Data Continuity Mission (LDCM, or Landsat 8 after launch), was launched in February 2013 carrying the next-generation Landsat thermal-imaging sensor.

The Landsat TM data is one of the most widely used satellite images for LST retrieving because of its high resolution (120 m) and free download availability from the website of US Geological Survey (USGS), which has one thermal infrared (TIR) band. This makes retrieving LST from a single band more difficult than from multiple thermal bands. In comparison, ASTER data has five thermal bands with a higher resolution (90 m), which may provide more promising potential for LST retrieval studies, although very few studies of LST retrieval from ASTER data are available as yet. Therefore, in this study, we applied the mono-window algorithm to the Landsat TM and the split-window algorithm to ASTER data for the analysis of its effect of urban heat island in the

Table 1 Remote-Sensing Satellite Sensors for measuring LST and their spatial resolutions

Remote-sensing sensor	Spatial resolution (meter)
ASTER	90
Landsat 3 MSS	240
Landsat 4, 5 TM	120
Landsat 7 ETM+	60

case study of Hong Kong. Although satellite data (e.g., Landsat TM and ASTER thermal bands data) can be applied to examine the distribution of urban heat islands in places such as Hong Kong, the method still needs to be refined with in situ measurements of LST in future studies. Among others, Ho et al. [9] used Landsat ETM+ to map urban temperature and compared the results to the local weather station results. This comparison shows the high accuracy of the satellite based observations.

Table 1 compares the spatial resolutions of the most frequently used passive sensors. Higher resolution thermal images could also be obtained from airborne images. Even though the airborne images provide very high-resolution information, it requires a very expensive procedure to fly over a large area to collect information. If the observation needs to be repeated with a certain time frequency, this procedure opens technical and financial challenges. Therefore, satellite sensors are highly preferred both by researchers and other land-observation institutions.

Further Supportive Data that Satellite Remote Sensing Can Offer

Besides providing measurements about wind speed/direction, air quality, and the heat islands, satellite sensors can make more measurements about earth which could support detailed research in the field of urban airflow analysis. For instance, multispectral satellite images provide an opportunity to calculate normalized difference vegetation index (NDVI) which tells about the vegetation greenness in the area. Some researchers have used NDVI as a major indicator of the urban climate. Experiments showed that during summer-time NDVI is negatively correlated with

surface temperature. However, previous studies have also shown a causal relationship between NDVI and LST that is subject to seasonal variation. Furthermore, the response of LST to NDVI varies among different land cover types. Some studies have demonstrated that LST has a stronger correlation with other parameters which could be obtained by using multispectral satellite image bands. For instance, the normalized difference built-up index (NDBI) or vegetation fraction than with NDVI showed higher correlation to the urban LST.

Impervious surfaces are found in urban and suburban landscapes and can be related to population density and urbanization. It is shown that reliable quantification of UHI could be achieved by analyzing the relationship between LST and impervious surface areas, which are again calculated by using multispectral bands of the satellite images.

Besides providing vegetation and built-up area indicator parameters from the multispectral band ratios, satellite images also offer the opportunity of automatic mapping of detailed 2D or 3D structures and for observing the traffic density as well. In a previous study, Sirmacek et al. [12] have proposed a 3D building reconstruction method compared to the rooftop shape prediction accuracies when different satellite sensors are used. Huang et al. [10] have studied impacts of different building roof shapes on the urban airflow. Their study shows that automatic building detection and rooftop modeling methods might provide important information to enhance the urban air quality and airflow research. Taubenboeck et al. [15] have studied extracting patterns from satellite images which can indicate social groups and income distribution within urban areas. These studies might also provide more discussion points to the urban air quality analysis research. Last but not least, there is probably a significant correlation between road usage and air quality in urban areas. Zheng et al. [23] used AI to estimate urban air quality. They trained their models with features coming from satellite data, which indicate distances to the road network, length of the roads, and meteorological data. Their study showed the significant correlation of the urban air quality with the extracted road features. Therefore, both the passive

sensor based information and active sensor based information might be important satellite sensor based sources to study the urban air quality.

Challenges and Open Problems

In the previous sections, we have seen that the application of machine-learning methods can be helpful in the study of urban air pollution. Two approaches were covered, that is, air-pollution models – which consisted on directly modeling air pollution – and flow-dynamics models, which focused on inferring the behavior of pollutants from the dynamics of the flow. Nevertheless, both approaches exhibit limitations and areas of improvement. On the one hand, the application of machine-learning algorithms to air-pollution models requires large amounts of data to provide accurate predictions. On the other hand, modeling the dynamics of the flow, despite not needing such databases, requires running numerical simulations which have an important computational cost. Furthermore, the inference of pollutant dispersion from the dynamics of the flow is not direct and thus requires a proper validation.

Additionally, another challenge affecting both approaches is the training of networks. It is known that the network training and the employed data usually have an important influence on the outcome of the system. In fact, the matter of training biases and, by extension, the biases of the system are important questions in all the fields where artificial intelligence is applied. The assessment of this problematic is relevant both from a technical perspective – since it heavily affects the outcome of the predictions produced by the network – and from an ethical point of view, since the outcome of those networks will typically influence policy-making decisions.

On the other hand, active sensors can help with direct measurements of the wind speed and direction. However, due to their low resolutions, their application areas are limited with sea and ocean monitoring. Their resolutions cannot help with adding valuable information to study urban airflow. On the other hand, passive sensors of the satellites provide measurements which can indirectly help with the urban airflow studies. They

can measure air quality, local heat, vegetation density and greenness. They can even provide detailed 3D building models and information about the road network. Even though these parameters are highly correlated with urban airflow, for precise mapping, still it is important to calibrate and enhance the satellite based information with in situ sensors. Budde et al. [2] introduced SmartAQnet called study which aims to combine multiple data resources to study the urban air quality in detail. They have combined IoT data such as weather sensors, in-situ air quality sensors, dust measurement devices, satellite images, and high-resolution drone images. However, due to the scalability issue of such a complex sensor fusion challenge, they have limited their analysis only with the Augsburg city of Germany. This study shows the big data collection and analysis challenges which still exist.

Another challenge is the validation of the health impacts of the urban airflow analysis. Even though the earlier studies showed a correlation with respiratory diseases, further research is still needed to completely pinpoint the urban-airflow aspects responsible for it. This will allow to develop better residence health estimation models, for increased urban sustainability.

References

- Brewer TL. Black carbon emissions and regulatory policies in transportation. *Energy Policy*. 2019;129: 1047–55. <https://doi.org/10.1016/j.enpol.2019.02.073>.
- Budde M, et al. SmartAQnet: remote and in-situ sensing of urban air quality. In: Proceedings of SPIE 10424, Remote Sensing of Clouds and the Atmosphere XXII, 104240C, vol. 1, no. 1. 2017. <https://doi.org/10.1117/12.228269>.
- Doreswamy HKS, Ibrahim Gad Yogesh KM. Forecasting air pollution particulate matter using machine learning regression models. *Procedia Comput Sci*. 2020;171:2057–66. <https://doi.org/10.1016/j.procs.2020.04.221>.
- European Commission. Urbanisation worldwide. https://ec.europa.eu/knowledge4policy/foresight/topic/continuing-urbanisation/urbanisation-world-wide_en
- Fujisada H, et al. ASTER DEM performance. *IEEE Trans Geosci Remote Sens*. 2005;43(12):2707–14. <https://doi.org/10.1109/TGRS.2005.847924>.
- Giri S, et al. Evaluating the impact of land uses on stream integrity using machine learning algorithms. *Sci Total Environ*. 2019;696(15):133858. <https://doi.org/10.1016/j.scitotenv.2019.133858>.
- Guastoni L, et al. Convolutional-network models to predict wall-bounded turbulence from wall quantities. *J. Fluid Mech*, To Appear 2021. arXiv preprint arXiv:2006.12483. 2020a.
- Guastoni L, et al. Prediction of wall-bounded turbulence from wall quantities using convolutional neural networks. *J Phys: Conf Ser*. 2020b;1522:012022.
- Ho HC, et al. A comparison of urban heat islands mapped using skin temperature, air temperature, and apparent temperature (Humidex), for the greater Vancouver area. *Sci Total Environ*. 2016;544(1):929–38. <https://doi.org/10.1016/j.scitotenv.2015.12.021>.
- Huang Y, et al. Impact of wedge-shaped roofs on airflow and pollutant dispersion inside urban street canyons. *Build Environ*. 2009;44(12):2335–47. <https://doi.org/10.1016/j.buildenv.2009.03.024>.
- Lelieveld J, et al. Cardiovascular disease burden from ambient air pollution in Europe reassessed using novel hazard ratio functions. *Eur Heart J*. 2019;40: 1590–6.
- Sirmacek B, et al. Performance evaluation for 3-D city model generation of six different DSMs from air- and spaceborne sensors. *IEEE J Select Topics Appl Earth Observ Remote Sens*. 2012;5(1):59–70. <https://doi.org/10.1109/JSTARS.2011.217839>.
- Srinivasan PA, et al. Predictions of turbulent shear flows using deep neural networks. *Phys Rev Fluids*. 2019;4:054603.
- Suleiman A, et al. Hybrid neural networks and boosted regression tree models for predicting roadside particulate matter. *Environ Model Assess*. 2016;21:731–50. <https://doi.org/10.1007/s10666-016-9507-5>.
- Taubenboeck H, et al. Integrating remote sensing and social science. *Joint Urban Remote Sens Event*. 2009;1 (1):1–7. <https://doi.org/10.1109/URS.2009.5137506>.
- Téllez-Rojo MM, et al. Children’s acute respiratory symptoms associated with PM2.5 estimates in two sequential representative surveys from the Mexico City Metropolitan Area. *Environ Res*. 2020;180(1): 108868. <https://doi.org/10.1016/j.envres.2019.108868>.
- Torres P, et al. The structure of urban flows. *Energies*. 2020;1–35. <https://doi.org/10.20944/preprints202009.0556.v1>.
- Vinuesa R, et al. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nat Commun*. 2020;11:233.
- Wang Z, et al. Acute health impacts of airborne particles estimated from satellite remote sensing. *Environ Int*. 2013;51(1):150–9. <https://doi.org/10.1016/j.envint.2012.10.011>.
- Wang A, et al. Potential of machine learning for prediction of traffic related air pollution. *Transp Res Part D: Transp Environ*. 2020;88:102599. <https://doi.org/10.1016/j.trd.2020.102599>.
- World Health Organization. Review of evidence on health aspects of air pollution. REVIHAAP Project, vol. 309. 2013. <http://www.euro.who.int/en/health-topics/environment-and-health/air-quality/>

- publications/2013/review-of-evidence-on-health-aspects-of-air-pollution-revihaap-project-final-technical-report
22. Xiao D, et al. A reduced order model for turbulent flows in the urban environment using machine learning. *Build Environ.* 2019;148:323–37. <https://doi.org/10.1016/j.buildenv.2018.10.035>.
23. Zheng Y, et al. U-air: when urban air quality inference meets big data. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD’13). Association for Computing Machinery, New York, vol. 1, no. 1. 2013. p. 1436–44. <https://doi.org/10.1145/2487575.2488188>.



AIM in Pharmacology and Drug Discovery

46

Hiroaki Iwata, Ryosuke Kojima, and Yasushi Okuno

Contents

Introduction	636
Ligand Screening and Pharmacology	636
Ligand-Based Approach	637
Structure-Based Approach	637
Chemical Genomics-Based Approach and Polypharmacology	638
ADME in Pharmacokinetics	638
Absorption	638
Distribution	638
Metabolism	639
Excretion	639
Data Source for Drug Discovery	639
Omics Data	640
Real-World Data	640
Summary and Future Implications	641
References	641

Abstract

Pharmaceutical products are researched and developed through many stages: target search, hit search, hit to lead, lead optimization, and preclinical and clinical trials, all of which have a complex, expensive, and time-

consuming process and a low overall success rate. Therefore, there is a need to reduce research and development costs by increasing the probability of success and improving process efficiency. One of the most promising approaches to this problem is the so-called *in silico* drug discovery, i.e., drug discovery using information technology such as artificial intelligence (AI) and molecular simulation. In this chapter, we describe the development and application of AI in drug discovery.

H. Iwata · R. Kojima · Y. Okuno (✉)
Department of Biomedical Data Intelligence, Graduate
School of Medicine, Kyoto University, Kyoto, Japan
e-mail: iwata.hiroaki.3r@kyoto-u.ac.jp;
kojima.ryosuke.8e@kyoto-u.ac.jp;
okuno.yasushi.4c@kyoto-u.ac.jp

Keywords

Chemistry · Chemoinformatics · Virtual screening · Omics data · Real-world data · Pharmacokinetics · ADME · Pharmaceutics · AlphaFold · Graph convolutional neural networks

Introduction

Drug discovery and development processes can be broadly divided into three categories: basic research, nonclinical studies, and clinical studies. Basic research begins with the search for target molecules, followed by the identification of hit compounds from a vast compound library and optimization of the hit compounds. Nonclinical studies are conducted in animals to test drug pharmacodynamics, *in vivo* dynamics, and adverse effects. Clinical trials can then confirm the safety (side effects) and efficacy in humans. Since drugs are developed through many stages, drug discovery is complex, expensive, and time-consuming and has a low overall success rate. Therefore, various AI technologies are being developed worldwide, such as prediction of target proteins based on available studies and omics information; protein 3D structure prediction; screening of hit compounds; automatic generation of lead chemical structures and prediction of synthetic pathways; prediction of absorption, distribution, metabolism, and excretion (ADME) of drugs in the body; prediction of side effects; and decision support in clinical trials. The main purpose of AI technologies is to reduce the probability of failure by making accurate predictions before the actual experiments and tests are conducted, and it is expected to optimize costs and duration of development by reducing the number of experiments and tests.

In recent years, *in silico* drug discovery is an actively implemented approach that uses computational methods such as molecular simulation and machine learning to improve the efficiency of the drug discovery process. This is due to the progress in AI technology and big data. AI technologies range from conventional machine learning, such as support vector machine (SVM) and random forest (RF), to deep learning, including deep neural networks

(DNNs), convolutional neural networks (CNNs), and graph convolutional networks (GCNs). For big data, various databases are available, such as PubChem/Bioassay [1], the world's largest bioactivity database published by the National Institutes of Health's National Center for Biotechnology Information (NIH/NCBI) in the United States; ChEMBL [2] and DrugBank [3], which contain information on the activities of compounds and target proteins; SIDER [4] and Food and Drug Administration Adverse Events Reporting System (FDA FAERS), which contain information on compounds and side effects; and the Kyoto Encyclopedia of Genes and Genomes (KEGG) [5], which is a database that integrates information on diseases, drugs, and intermolecular networks such as metabolism and signal transduction.

Here, we focus on three topics that are important in the application of AI in drug discovery: (1) ligand screening and pharmacology; (2) absorption, distribution, metabolism, and excretion (ADME) in pharmacokinetics; and (3) data source for drug discovery.

Ligand Screening and Pharmacology

There are various difficulties in drug development, but one of the most challenging stages is the initial selection of compounds from the large pharmacological space. Theoretically, there are more than 10^{60} variations of molecular structures and more than 100,000 variations of target molecules to which a compound binds, even when limited to proteins, which are the most typical targets. Therefore, to efficiently search for the best compound-protein combination, advanced methods such as machine learning, especially deep learning, are necessary in addition to conventional ligand docking simulations. Various machine-learning methods have been applied to predict active compounds. Since Hinton's group won the state-of-the-art in the Kaggle competition held by Merck, a major pharmaceutical company, in 2012 for a deep learning method using a multitask deep neural network [6, 7], applications using deep learning have been rapidly advancing [8]. In this section, we will explain three computational approaches for the prediction of

compound-protein interactions in pharmacology: ligand-based, structure-based, and chemical genomics-based [9, 10].

Ligand-Based Approach

The ligand-based approach was developed based on the idea that compounds with similar chemical structures have similar pharmacological activities [11]. Therefore, the model inputs the chemical structure information of a candidate compound and then outputs the presence or absence of pharmacological activity such as bioactivity for disease cells and binding to target proteins. For example, in a quantitative structure-activity relationship-like (QSAR) approach, a machine-learning method is used to compare a candidate ligand with all known ligands and infer its binding ability. However, this approach has some drawbacks. Since protein information is not used for learning, the novel interactions that can be predicted are limited to the relationships between known ligands and protein families. In addition, this method requires the construction of a model for each target protein, and while it is possible to build an accurate model for proteins with abundantly known ligands, it becomes difficult to accurately predict binding active compounds conversely [12]. Multitask learning has been proposed as a technique to overcome this problem and has been recently used frequently [7]. Multitask learning is a technique for simultaneously learning pharmacological activity data for each of the multiple cell phenotypes and target proteins, which can be achieved with a network configuration that shares some of the hidden layers. Multitask learning is expected to improve the overall process accuracy because it can learn the features of active compounds that are common among multiple phenotypes/proteins; even when the scale of binding information for each phenotype/protein is small, they can be efficiently learned in a way that complements each other's data volume.

Structure-Based Approach

The structure-based approach uses the 3D structures of proteins and compounds to perform docking simulations on the pocket sites of target

proteins to predict whether they will bind or not. Unlike the ligand-based approach, it does not require known active compounds but allows the search for active compounds with new scaffolds [13]. However, this approach also has a crucial limitation. In fact, defining the 3D structure of a target protein is difficult as many target proteins have not been elucidated. In these cases, a structure-based approach cannot be applied.

Predicting the 3D structures of a protein from sequence information, although mainly observed for decades during the critical assessment of structure prediction (CASP) competition, is an extremely challenging issue. Recently, AlphaFold, a combination of deep learning and simulation techniques, has attracted much attention because of its superiority over existing methods [14]. In particular, in the CASP14 experiment, AlphaFold2 outperformed all its competitors and became a hot topic [15].

Various methods have been developed for the prediction of the 3D structure of proteins, including template-based methods and relaxation, using profiles and fragments. Many of the high-ranked systems used in the competition consist of modules to search in databases and contact map estimation and machine learning, including deep learning [16, 17]. In addition, it is often used in combination with a simulation module that uses an empirically based force field because it is difficult to perform a complete simulation based on the first principle [18, 19]. Although deep learning was used for the contact map estimation and multiple sequence alignment modules for 3D structure prediction before the appearance of AlphaFold, its first success was produced by expressing the uncertainty of contact map estimation and constructing a deep learning system with the simulation. Furthermore, a higher performance in AlphaFold2 was achieved by using the end-to-end architecture instead of using the intermediate step for the contact map estimation. In general, many deep learning applications have successfully replaced the traditional manual and sophisticated pipeline of multiple modules with a differentiable end-to-end architecture based on traditional processes. With the success of AlphaFold2, such an approach may become an important trend in other drug discovery issues.

Chemical Genomics-Based Approach and Polypharmacology

The chemical genomics-based approach predicts potential interactions by integrating the chemical space of drugs and the biological space of proteins, using both chemical structure similarity of compounds and sequence similarity of proteins as indicators [10]. This approach is expected to reveal complicated multiple interactions between compounds and proteins, that is, polypharmacology. Furthermore, the chemical genomics-based approach overcomes the disadvantages of ligand-based and structure-based approaches for various reasons. First, the fundamental advantage is the availability of a wide range of publicly available biological data. For example, the chemical structures and protein sequences of thousands of compounds have already been compiled in public databases [1, 2, 20]. This information can be used to predict the interactions [21]. Since the prediction is based on both chemical structure similarity of compounds and sequence similarity of proteins, it is more accurate than the ligand-based methods that use only compound information [22]. Second, the chemical genomics-based approach is more versatile than the ligand-based and structure-based approaches because it can be applied to target proteins for which there are many assay data or of which the 3D structure is not yet known, even if there is no known active compound information of the target protein itself. Third, the chemical genomics-based approach is the mainstream of recent programs that predict compound-protein interactions. For example, there is a multimodal method in which the chemical structure of a compound is learned by Graph CNNs and the protein sequence information is learned by CNNs, and they are combined in the middle of the network [23]. Another example is a multimodal method where the compound fingerprint and protein are trained with multilayer perceptron and CNN [24], respectively, and the compound and protein are trained with CNN and long short-term memory (LSTM) [25].

ADME in Pharmacokinetics

Pharmacokinetic studies are involved in the entire process, from compound seed discovery to clinical trials, and play an important role in new drug development [26, 27]. It was reported that pharmacokinetics was the reason for about 40% of dropouts in drug discovery and development until 1985 [28]. Subsequently, due to improvements in in vitro absorption, distribution, metabolism, and excretion (ADME) assays, the rate decreased approximately from 40% to 10% between 1991 and 2000 [29, 30]. One strategy to further improve the success rate of clinical trials is to estimate human clinical doses that exhibit optimal drug efficacy profiles. This requires accurate prediction of human pharmacokinetic parameters from nonclinical data before moving on to human clinical trials [31]. The development of high-throughput in vitro screening technologies has led to the accumulation of high-quality high-throughput ADME data, and in silico prediction of ADME properties using machine-learning methods has also gained significant momentum [32–35]. We introduce the modeling of ADME, which has been reported in recent years.

Absorption

Human intestinal absorption (HIA) is the process by which a drug is transferred from the site of administration to the vascular system. It is a process that determines bioavailability and has a significant impact on the efficacy and safety of drugs. Some studies have used SVM, artificial neural network (ANN), k-nearest neighbor (k-NN), probabilistic neural network (PNN), partial least squares (PLS), and linear discriminant analysis (LDA) to predict HIA and finally reported that the SVM algorithm with a kinetic basis function kernel had the best accuracy [36].

Distribution

Distribution refers to the spread of a drug throughout the vascular network of the body after it enters

the body and its arrival in the target tissue or organ. Prediction of the volume of distribution from nonclinical data has been relatively successful because it is largely determined by the physical properties of the drug, such as protein binding and membrane permeability. For example, QSAR studies have been performed to predict four human pharmacokinetic parameters: volume of distribution at steady state (VD_{ss}), clearance (CL), terminal half-life ($t_{1/2}$), and unbound fraction in plasma (fu). The method is based on feature selection using molecular descriptors of MOE, moldred, and PaDEL, and prediction models are constructed using several conventional machine-learning methods. These models demonstrated excellent stability and prediction ability [37]. A new random forest model has also been proposed, which depends only on six descriptors: MoKa-derived maximum basic pKa (v. 2.6.6, Molecular Discovery, Ltd., UK), minimum acidic pKa, molecular weight, distribution coefficients at pH 7.4 ($\log D_{7.4}$), number of aromatic atoms, and the presence or absence of a sulfur atom. An evaluation of this model using an independent test set showed that it performed as well as the previous benchmark model, which relied on a much larger number of descriptors [38].

Metabolism

Drug metabolism refers to the chemical structural changes that a drug undergoes once it has been delivered. Both experimental and computational approaches have been developed to investigate the metabolism and fate of drugs [39]. Many advances in the prediction of drug metabolism using in silico approaches have been made as part of the drug discovery effort, many of which have been reviewed. Commercial software includes Meteor Nexus, a knowledge-based expert system that predicts the metabolism of a compound from its structure; MetabolExpert (CompuDrug, Bal Harbor, FL, USA) (<http://www.computdrug.com/metabolexpert>); ADMET Predictor (Simulation Plus, Lancaster, CA, USA) (<https://www.simulations-plus.com/software/admetpredictor/metabolism>); and rule-based

expert systems. Free software includes XenoSite [40], a CYP SoM prediction model based on neural networks that improves rank-based stochastic pooling (RSP) in multiple ways, and SMARTCyp [41], which predicts the metabolic sites of compounds via cytochrome P450.

Excretion

Excretion is the process by which drugs are removed from the body involving the liver and kidneys. An animal scale-up was proposed to attempt an empirical prediction of the pharmacokinetic parameters of total body clearance, which is the ability to process drugs in humans. This method is based on the power law of the animal's weight. The error in predicting the clearance value using this method is on average twice as large as the error, but it has not been validated on different datasets. Other methods have been proposed to apply machine-learning methods with fingerprints of chemical structures, physicochemical parameters, and animal clearances as explanatory variables [42, 43]. In addition, some studies have shown excellent prediction results using a multimodal method that uses descriptors calculated from chemical structures (or the graph itself) and clearance value of rats as explanatory variables for predicting those of humans [44].

Data Source for Drug Discovery

For developing an AI, it is extremely important to acquire relevant data. Recent advances in genomics, sequencing, and high-throughput technologies have generated a large amount of diverse data for drug discovery. To use these datasets to search for highly successful target molecules, it is important to elucidate the mechanism of action. In this section, we introduce research using omics and real-world data as big data. The term "omics data" refers to data from the genome, which is the total genetic information; epigenome, which is the information on the modification of the genome; transcriptome, which is the information on all transcripts; proteome, which is the information

on the proteins produced; and metabolome, which is the information on the metabolites produced by the interaction of proteins. On the contrary, real-world data refer to data collected from sources outside the traditional research environment, such as electronic health records (EHRs), administrative claims, and billing data [45].

Omics Data

In drug discovery, the mechanisms of the target disease and of action and side effects of new drugs at the cellular level can be explained by measuring and analyzing omics data. Single nucleotide polymorphisms (SNPs) and copy number polymorphisms (CNVs) have been widely used to characterize diseases. SNPs and CNVs can be identified from genome-wide association studies (GWAS) and whole-genome sequencing approaches and have been comprehensively investigated by next-generation sequencing and used to identify driver genes responsible for selective pathogenic cell proliferation. In addition, with the development of microarray technology, gene expression is measured and used to understand the mechanisms of disease. Recently, expression data of mRNA and ncRNA by RNA-Seq have become available [46].

The analysis of drug discovery and biological data using omics data and information technology (IT) has been conducted for a long time, for example, SNP and GWAS analyses for genomic data, differential expression analysis, and cluster analysis of transcriptome data. With the development of IT, it is now possible to prepare a large amount of omics data and systematically and comprehensively analyze the hypothesized mechanism of action of complex multiple factors. Specifically, this includes gene network estimation and analysis of transcriptome data and protein-protein interaction (PPI) network analysis. These are based on single-omics data, but further development of multiple-omics analysis techniques is currently pursued [47].

The relationships among omics data can be viewed as a multilayered network of omics

species. The lowest layer is the genome and epigenome and above the transcriptome, proteome, and metabolome networks. The interaction of each layer cannot be explained by analyzing single-omics species. If the relationships among biomolecules in the cell are considered as such a multilayered network, then the mechanism of this network is explained using multi-omics data. Information on this network has been collected and edited from the literature and other sources and stored in databases such as KEGG but is limited to a small portion of networks and is not exhaustive. Therefore, it is important to predict and estimate these networks from a large amount of omics data and to compare and analyze them with literature-based knowledge to clarify the mechanisms.

Real-World Data

In recent years, the FDA has been promoting the use of real-world data (RWD) in drug discovery and development [48], which has gained momentum [49, 50]. A recent review [49] on the use of AI in real-world data surveyed 20 studies that used RWD to facilitate drug discovery and clinical research. Sixteen of these studies identified and validated novel phenotypes, disease markers, and biomarkers for patient identification and stratification. Another review [50] identified and surveyed 65 studies by screening titles, abstracts, and full texts. The number of studies using RWD and AI methods in the drug development process, especially in the clinical or post-marketing phase, is therefore on the rise. Most of the data sources were EHRs, especially unstructured clinical notes. The most commonly used AI method was NLP, which is a trans-based model for clinical concept extraction such as bidirectional encoder representations from transformers (BERT) model [51, 52], because it deals with unstructured data. The studies also focused on adverse event detection, optimization of clinical trial recruitment, and drug repositioning. First, in a study on adverse event detection, an NLP system was developed to identify the adverse drug administration by extracting

information from clinical notes. For example, a relationship extraction system using bidirectional long short-term memory (BiLSTM) networks, conditional random fields (CRF) [53], and a recurrent neural network (RNN) model using a LSTM strategy [54] was used to perform clinical name entity recognition (NER) to extract and classify relationships between drugs and drug-related entities. Next, to optimize clinical trial recruitment, most of the studies used clinical notes from EHR data, using information extracted by NLP to identify the populations eligible for clinical trials [55]. Finally, in a study of drug repositioning from clinical data, the use of EHRs and informatics methods was used to suggest that the use of metformin, a diabetes drug, in oncology was associated with lower cancer mortality [56]. However, current research using RWD often relies on a single EHR system and biases the analysis of specific drugs and indications. In addition, the complex and heterogeneous nature of clinical documentation and data quality limits drug repositioning studies [57].

Summary and Future Implications

In this chapter, we present the application of AI and related databases to ligand screening, pharmacology, ADME, and pharmacokinetics in drug discovery processes. Since the actual development process comprises more stages, drug discovery is complex, expensive, and time-consuming and has a low overall success rate. AI technologies are expected to dramatically decrease the cost and time of development by reducing the number of experiments and tests. However, many of the methods introduced here are still under research and development with only a few put to practical use in actual drug discovery. However, considering the limitations of the black box and the quality and quantity of training data needed, various crucial technical issues remain to be solved in AI technologies. The rapid development of AI technologies and the expansion of problem-solving methods will continue to drive the application of AI in the drug discovery field.

References

- Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* 2009;37:W623–33.
- Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 2012;40:D1100–7.
- Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 2018;46:D1074–D82.
- Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. *Nucleic Acids Res.* 2016;44:D1075–9.
- Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000;28:27–30.
- Markoff J. Scientists see promise in deep-learning programs. *New York Times.* 2012;23.
- Dahl GE, Jaitly N, Salakhutdinov R. Multi-task neural networks for QSAR predictions. *arXiv preprint arXiv:14061231.* 2014.
- Gawehn E, Hiss JA, Schneider G. Deep learning in drug discovery. *Mol Inform.* 2016;35:3–14.
- Sachdev K, Gupta MK. A comprehensive review of feature based methods for drug target interaction prediction. *J Biomed Inform.* 2019;93:103159.
- Ezzat A, Wu M, Li XL, Kwoh CK. Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey. *Brief Bioinform.* 2019;20:1337–57.
- Johnson MA, Maggiora GM. Concepts and applications of molecular similarity. Wiley; 1990.
- Jacob L, Vert JP. Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics.* 2008;24:2149–56.
- Jimenez J, Skalic M, Martinez-Rosell G, De Fabritiis G. KDEEP: protein-ligand absolute binding affinity prediction via 3D-convolutional neural networks. *J Chem Inf Model.* 2018;58:287–96.
- Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins.* 2019;87:1141–8.
- Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Tunyasuvunakool K, et al. High accuracy protein structure prediction using deep learning. Fourteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstract Book). 2020;22:24.
- Zheng W, Li Y, Zhang C, Pearce R, Mortua SM, Zhang Y. Deep-learning contact-map guided protein structure prediction in CASP13. *Proteins.* 2019;87: 1149–64.
- Li Y, Zhang C, Bell EW, Yu DJ, Zhang Y. Ensembling multiple raw coevolutionary features with deep

- residual neural networks for contact-map prediction in CASP13. *Proteins*. 2019;87:1082–91.
18. Park H, Kim DE, Ovchinnikov S, Baker D, DiMaio F. Automatic structure prediction of oligomeric assemblies using Robetta in CASP12. *Proteins*. 2018;86(Suppl 1):283–91.
19. Hong SH, Joung I, Flores-Canales JC, Manavalan B, Cheng Q, Heo S, et al. Protein structure modeling and refinement by global optimization in CASP12. *Proteins*. 2018;86(Suppl 1):122–35.
20. UniProt C. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*. 2019;47:D506–D15.
21. Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*. 2008;24:i232–40.
22. Yabuuchi H, Niijima S, Takematsu H, Ida T, Hirokawa T, Hara T, et al. Analysis of multiple compound-protein interactions reveals novel bioactive molecules. *Mol Syst Biol*. 2011;7:472.
23. Tsubaki M, Tomii K, Sese J. Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*. 2019;35:309–18.
24. Lee I, Keum J, Nam H. DeepConv-DTI: prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS Comput Biol*. 2019;15:e1007129.
25. Abbasi K, Razzaghi P, Poso A, Amanlou M, Ghasemi JB, Masoudi-Nejad A. DeepCDA: deep cross-domain compound-protein affinity prediction through LSTM and convolutional neural networks. *Bioinformatics*. 2020;36:4633–42.
26. Ballard P, Brassil P, Bui KH, Dolgos H, Petersson C, Tunek A, et al. The right compound in the right assay at the right time: an integrated discovery DMPK strategy. *Drug Metab Rev*. 2012;44:224–52.
27. Ferreira LLG, Andricopulo AD. ADMET modeling approaches in drug discovery. *Drug Discov Today*. 2019;24:1157–65.
28. Prentis R, Lis Y, Walker S. Pharmaceutical innovation by the seven UK-owned pharmaceutical companies (1964–1985). *Br J Clin Pharmacol*. 1988;25:387–96.
29. MacCoss M, Baillie TA. Organic chemistry in drug discovery. *Science*. 2004;303:1810–3.
30. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov*. 2004;3:711.
31. Andrade EL, Bento AF, Cavalli J, Oliveira SK, Schwanke RC, Siqueira JM, et al. Non-clinical studies in the process of new drug development – Part II: Good laboratory practice, metabolism, pharmacokinetics, safety and dose translation to clinical studies. *Braz J Med Biol Res*. 2016;49:e5646.
32. Shou WZ. Current status and future directions of high-throughput ADME screening in drug discovery. *J Pharm Anal*. 2020;10:201–8.
33. Maltarollo VG, Gertrudes JC, Oliveira PR, Honorio KM. Applying machine learning techniques for ADME-Tox prediction: a review. *Expert Opin Drug Metab Toxicol*. 2015;11:259–71.
34. Wenzel J, Matter H, Schmidt F. Predictive multitask deep neural network models for ADME-Tox properties: learning from large data sets. *J Chem Inf Model*. 2019;59:1253–68.
35. Chen PC, Liu Y, Peng L. How to develop machine learning models for healthcare. *Nat Mater*. 2019;18: 410–4.
36. Kumar R, Sharma A, Siddiqui MH, Tiwari RK. Prediction of human intestinal absorption of compounds using artificial intelligence techniques. *Curr Drug Discov Technol*. 2017;14:244–54.
37. Wang Y, Liu H, Fan Y, Chen X, Yang Y, Zhu L, et al. In silico prediction of human intravenous pharmacokinetic parameters with improved accuracy. *J Chem Inf Model*. 2019;59:3968–80.
38. Lombardo F, Bentzien J, Berellini G, Muegge I. In silico models of human PK parameters. Prediction of volume of distribution using an extensive data set and a reduced number of parameters. *J Pharm Sci*. 2021;110: 500–9.
39. Kirchmair J, Goller AH, Lang D, Kunze J, Testa B, Wilson ID, et al. Predicting drug metabolism: experiment and/or computation? *Nat Rev Drug Discov*. 2015;14:387–404.
40. Zaretzki J, Matlock M, Swamidass SJ. XenoSite: accurately predicting CYP-mediated sites of metabolism with neural networks. *J Chem Inf Model*. 2013;53: 3373–83.
41. Olsen L, Montefiori M, Tran KP, Jorgensen FS. SMARTCyp 3.0: enhanced cytochrome P450 site-of-metabolism prediction server. *Bioinformatics*. 2019;35:3174–5.
42. Wajima T, Fukumura K, Yano Y, Oguma T. Prediction of human clearance from animal data and molecular structural parameters using multivariate regression analysis. *J Pharm Sci*. 2002;91:2489–99.
43. Huang W, Geng L, Deng R, Lu S, Ma G, Yu J, et al. Prediction of human clearance based on animal data and molecular properties. *Chem Biol Drug Des*. 2015;86:990–7.
44. Iwata H, Matsuo T, Mamada H, Motomura T, Matsushita M, Fujiwara T, et al. Prediction of total drug clearance in humans using animal data: proposal of a multimodal learning method based on deep learning. *J Pharm Sci*. 2021;110:1834.
45. Makady A, de Boer A, Hillege H, Klungel O, Goettsch W. What is real-world data? A review of definitions based on literature and stakeholder interviews. *Value Health*. 2017;20:858–65.
46. Chen B, Butte AJ. Leveraging big data to transform target selection and drug discovery. *Clin Pharmacol Ther*. 2016;99:285–97.
47. Arima C, Kajino T, Tamada Y, Imoto S, Shimada Y, Nakatohchi M, et al. Lung adenocarcinoma subtypes definable by lung development-related miRNA expression profiles in association with clinicopathologic features. *Carcinogenesis*. 2014;35:2224–31.

48. FDA U. Use of real-world evidence to support regulatory decision-making for medical devices. Guidance for Industry and Food and Drug Administration Staff. 2017.
49. Singh G, Schulthess D, Hughes N, Vannieuwenhuyse B, Kalra D. Real world big data for clinical research and drug development. *Drug Discov Today*. 2018;23:652–60.
50. Chen Z, Liu X, Hogan W, Shenkman E, Bian J. Applications of artificial intelligence in drug development using real-world data. *Drug Discov Today*. 2020;26:1256.
51. Yang X, Bian J, Hogan WR, Wu Y. Clinical concept extraction using transformers. *J Am Med Inform Assoc*. 2020;27:1935–42.
52. Fu S, Chen D, He H, Liu S, Moon S, Peterson KJ, et al. Clinical concept extraction: a methodology review. *J Biomed Inform*. 2020;109:103526.
53. Christopoulou F, Tran TT, Sahu SK, Miwa M, Ananiadou S. Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. *J Am Med Inform Assoc*. 2020;27:39–46.
54. Yang X, Bian J, Gong Y, Hogan WR, Wu Y. MADEx: a system for detecting medications, adverse drug events, and their relations from clinical notes. *Drug Saf*. 2019;42:123–33.
55. Opella SJ. Structure determination of membrane proteins by nuclear magnetic resonance spectroscopy. *Annu Rev Anal Chem (Palo Alto, Calif)*. 2013;6: 305–28.
56. Xu H, Aldrich MC, Chen Q, Liu H, Peterson NB, Dai Q, et al. Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. *J Am Med Inform Assoc*. 2015;22:179–91.
57. Xu H, Li J, Jiang X, Chen Q. Electronic health records for drug repurposing: current status, challenges, and future directions. *Clin Pharmacol Ther*. 2020;107: 712–4.



Clinical Evaluation of AI in Medicine

47

Xiaoxuan Liu, Gagandeep Sachdeva, Hussein Ibrahim,
Maria Charalambides, and Alastair K. Denniston

Contents

Clinical AI Systems Require Robust Clinical Evaluation	646
Randomized Controlled Trials of Clinical AI Systems	647
AI Systems for Disease Diagnosis	647
AI Systems for Disease Prediction	655
AI Systems for Adjusting Therapeutic Treatment	656
Reporting Standards for AI Clinical Trials	656
New Reporting Guidelines for AI Trials to Reflect the New Epoch	657
Future Challenges	658
References	658

X. Liu (✉)

University Hospitals Birmingham NHS Foundation Trust,
Birmingham, UK

Centre for Regulatory Science and Innovation,
Birmingham Health Partners, University of Birmingham,
Birmingham, UK

College of Medical and Dental Sciences, University of
Birmingham, Birmingham, UK
e-mail: x.liu.8@bham.ac.uk

G. Sachdeva · M. Charalambides

College of Medical and Dental Sciences, University of
Birmingham, Birmingham, UK
e-mail: GSS763@student.bham.ac.uk;
MXC578@student.bham.ac.uk

H. Ibrahim

University Hospitals Birmingham NHS Foundation Trust,
Birmingham, UK

Centre for Regulatory Science and Innovation,
Birmingham Health Partners, University of Birmingham,
Birmingham, UK

e-mail: h.ibrahim.1@bham.ac.uk

A. K. Denniston

University Hospitals Birmingham NHS Foundation Trust,
Birmingham, UK

Centre for Regulatory Science and Innovation,
Birmingham Health Partners, University of Birmingham,
Birmingham, UK

College of Medical and Dental Sciences, University of
Birmingham, Birmingham, UK

Health Data Research UK, London, UK

National Institute of Health Research Biomedical Research
Centre for Ophthalmology, Moorfields Hospital London
NHS Foundation Trust and University College London,
Institute of Ophthalmology, London, UK
e-mail: A.Denniston@bham.ac.uk

Abstract

Clinical evaluation provides the necessary evidence that healthcare interventions are safe, effective, and likely to bring benefit to patients and healthcare system. While a large volume of evidence for in silico performance of AI systems has been published in recent years, prospective clinical trials are only just beginning to emerge. This chapter presents a summary of prospective randomized trials published in AI health interventions to date and discusses new AI-specific reporting standards for clinical evaluations.

Clinical AI Systems Require Robust Clinical Evaluation

The number of publications involving AI methods in biology has grown exponentially in recent years, with over 21,000 papers published in 2020 alone [1], and the global AIM market has grown from \$600 million in 2014 to \$6.6 billion in 2021 [2]. A 2021 study reports 222 AI/ML-based devices approved by regulatory authorities in the USA and 240 devices in Europe to date [3].

Despite this, the widespread routine use of AI/ML-based medical devices is patchy, and concerns over their true clinical effectiveness remain. One obstacle preventing innovative AI technologies from being routinely implemented in clinical practice is a lack of evidence that they remain robust in real-world clinical environments and lead to demonstrable improvement in patient outcomes. Specifically, there has been a general lack of evidence from prospective, comparative clinical trials which focus on health outcomes, rather than measures of performance (such as accuracy) which are several degrees removed from measures of effectiveness.

Most existing evidence for AI healthcare interventions so far comes from retrospective, observational studies. The evidence, as demonstrated in two systematic reviews of AI medical imaging algorithms [4, 5], is promising but insufficient to indicate suitable immediate implementation in clinical practice. This is because these studies

are formulated to ask whether AI healthcare systems do what they were *designed* to do, not what they were *intended* to do. That is to say, they ask whether AI healthcare systems *predict, identify, or produce* a state or experience *reliably and accurately*, not whether they *safely and effectively* predict, identify, or produce a *clinically meaningful* state or experience, *when used in a specific population and setting*.

Evidence that an AI healthcare intervention is safe, clinically effective, and cost-effective is an essential prerequisite to implementation in clinical practice, if the ultimate goal is to bring benefits to patients and society. It is not enough to show that a diagnostic AI algorithm reliably and accurately detects disease from imaging data in a computational context. It also needs to be shown that when applied in a clinical setting, it improves patient outcomes and does so at least as well as healthcare professionals, without causing additional complications and/or harm.

The strongest evidence for the safety, clinical effectiveness, and cost-effectiveness of an AI intervention requires evaluation in the context of well-designed, well-delivered, and well-reported prospective, comparative clinical trials [6]. This is because two fundamental characteristics of clinical trials – randomization and a control arm – allow us to make causal inferences about interventions and their outcomes [6].

Even seemingly benign interventions, such as the example of an acute kidney injury alert system, have been shown to unexpectedly result in patient harm in randomized trials [7]. Findings such as this one, although arising from unclear mechanisms, could not have been discoverable without evaluation through prospective comparative trials. It therefore emphasizes the importance of evaluating AI healthcare interventions in their intended clinical setting, using health outcomes as the primary endpoint, and actively investigating potentially deleterious consequences. Crucially, the evidence from these studies should, like all trials, be reported to the highest standards, to ensure transparency and assessment of potential biases and enable the study to be appraised and the evidence evaluated. Only this way will stakeholders be able to establish which AI healthcare interventions provide real value.

Stakeholders need to know which claims made about AI systems are substantiated and which claims are not: patients, the public, healthcare professionals, healthcare providers, and medicine and medical device regulators need to know which AI healthcare systems are safe and clinically effective; and investors, payers, politicians, and health policy-makers need to know how cost-effective these AI healthcare systems are. The most rigorous and robust way to answer these questions is through prospective, comparative clinical trials. Therefore, although most AI healthcare interventions have not yet been evaluated in prospective, randomized clinical trials, this will be an area of rapid expansion moving forward, as the field of AIM continues to mature.

Randomized Controlled Trials of Clinical AI Systems

At the time of writing, while adoption of clinical AI systems remains in its infancy, there is a growing recognition that AI systems require evaluation in prospective, real-world, clinical settings. There is also an emphasis on such evaluations paying due consideration to the effect of AI systems on the clinical pathway, the representativeness of the population and setting, and importantly the relative benefits of such interventions on patient outcomes.

As of 2021, a comparatively small number of prospective clinical trials for AI systems have been conducted and published. We identified 14 completed and published trials as of December 2020 through the [ClinicalTrials.gov](#) registry, including AI interventions for diagnosis and disease prediction, AI as therapeutic interventions, and AI for streamlining or increasing the efficiency of clinical tasks. Of note, several randomized controlled trials (RCTs) in the field of colonic endoscopy have already been conducted, evaluating the effectiveness of AI for detecting colonic adenomas. This specific use case is by far the most mature in evidence generation across the field of clinical AI.

In this section, we present a selection of illustrative examples of published AI clinical trials as of 2021 (summarized in Table 1).

AI Systems for Disease Diagnosis

AI Systems for Detection of Colonic Adenoma During Endoscopy

Perhaps the most mature area of prospective trials is AI for detection of colonic adenoma. This is likely due to the real-time nature of AI systems applied to endoscopy, which would be otherwise difficult to evaluate in a simulated or observational study.

The adenoma detection rate (ADR) is regarded as the primary indicator of quality during colonoscopy, and an increased ADR through conventional clinician-led endoscopy has been shown to reduce the incidence and mortality of colorectal cancer [21, 22]. More recently, several deep learning models have been shown to yield higher ADRs compared to conventional endoscopy. These studies are summarized below.

A prospective trial of a real-time polyp detection system (Shanghai Wision AI Co., Ltd.) studied 1058 patients, randomized either to an AI-assisted polyp detection ($n = 522$) or to routine colonoscopy without AI ($n = 536$) [12]. In the intervention arm, the AI system was activated during the withdrawal phase of colonoscopy, providing real-time alerts with a hollow blue tracing box and sound alarm upon detection of a suspected polyp. A statistically superior ADR was reported with the AI system, compared to routine colonoscopy (29.1% vs 20.3%, $p < 0.001$). The mean number of adenomas detected per patient was significantly greater with AI (0.53 vs 0.31, $p < 0.001$), with superior detection of diminutive and hyperplastic polyps (185 vs 102 and 114 vs 52, respectively, $p < 0.001$). No significant difference was reported for the detection of larger adenomas ($p = 0.075$). The trial presented promising results and advocated for further research into determining cost-benefit ratios for AI-assisted colonoscopy.

The CADe-DB trial was a double-blinded randomized single-center study that compared ADRs of the intervention arm ($n = 484$) of colonoscopy with computer-aided detection (CADe) to the control arm ($n = 478$) of colonoscopy with a sham system [9]. Patients aged 18–75 presenting to the endoscopy center were consecutively enrolled.

Table 1 Published clinical trials evaluating clinical AI systems

Publication	Disease Context	Population	Intervention	Control	Outcome	Summary of results
Wang et al. [12]	Colonic adenoma Diagnosis	Consecutive enrollment of patients undergoing colonoscopy at the endoscopy center of the Sichuan Provincial People's Hospital, China (Sep. 2017–Feb. 2018)	A deep learning system for real-time automated polyp detection system ($n = 522$)	Standard colonoscopy, in the absence of the automated system ($n = 536$)	ADR	The AI system significantly improved ADR (29.1% vs 20.3%, $p < 0.001$) and the mean number of adenomas per patient (0.53 vs 0.31, $p < 0.001$), as compared to the control arm. This was reported as being due to greater detection of small adenomas (185 vs 102, $p < 0.001$) and hyperplastic polyps (114 vs 52, $p < 0.001$) with the AI system
Wang et al. [9]	Colonic adenoma	Consecutive enrollment of patients (aged 18–75 years) presenting for diagnostic or screening colonoscopy at the endoscopy center to Catang branch hospital of Sichuan Provincial Hospital, China (Sept. 2018–Jan. 2019)	Computer-aided detection (CADe) deep learning automatic polyp detection system ($n = 484$)	Shan system, simulating alerts on polyp-like non-polyp structures ($n = 478$)	Adenoma detection rate (ADR)	Significantly greater ADR in the CADe arm, as compared to the sham control arm (odds ratio 1.36, $p = 0.030$) One or more adenomas detected: 165 of 484 patients (34% in CADe), 132 of 478 in sham control (28%)

Su et al. [10]	Colonic adenoma	Consecutive enrollment of patients presenting for routine colonoscopy screening to Qilu Hospital, Shandong University (Oct. 2018–May 2019)	An automated quality control system (AQCS) to optimize colonoscopy examination ($n = 308$) The system provides withdrawal phase supervision, timer display, steady withdrawal speed prompts, audio alerts for inadequate exposure of bowel segments, and automated alert boxes upon polyp detection	Routine, non-AQCS-assisted colonoscopy ($n = 315$)	ADR, mean number of adenomas detected per procedure, adequate bowel preparation rate, withdrawal time	The AI system (AQCS) significantly improved adenoma detection rate (0.289 vs 0.165, $p < 0.001$), mean number of adenomas detected per procedure (0.367 vs 0.178, $p < 0.001$), adequate bowel preparation (87% vs 80%, $p = 0.023$), and withdrawal time (7.03 min vs 5.68 min, $p < 0.001$), compared to the control
Gong et al. [11]	Colonic adenoma	Consecutive enrollment of patients (aged 18–75 years) undergoing colonoscopy at the endoscopy center in Renmin Hospital, Wuhan University (Jun. 2019–Sep. 2019)	<i>ENDOANGEL</i> convolutional neural network system for adenoma detection ($n = 355$)	Routine, non- <i>ENDOANGEL</i> -assisted colonoscopy ($n = 349$)	ADR	<i>ENDOANGEL</i> significantly improved ADR, as compared to the control arm (odds ratio 2.30, $p = 0.001$) One or more adenomas detected: 58 of 355 patients (16%) in <i>ENDOANGEL</i> , 27 of 459 patients in control arm (8%)
Lui et al. [8]	Colonic adenoma	Consecutive enrollment of patients (aged 40 and older) scheduled for outpatient	Region-based fully connected convolutional neural network (R-FCN): a two-stage	Primary endoscopist performed routine colonoscopy (blinded to AI system) ($n = 52$) ^a	Number of adenomas detected	The AI system detected at least 1 missed adenoma in 14 patients and

(continued)

Table 1 (continued)

Publication	Disease Context	Population	Intervention	Control	Outcome	Summary of results
		colonoscopy at Queen Mary Hospital, University of Hong Kong (Jan. 2020–Feb. 2020)	convolutional neural network which [1] identifies regions of interest and [2] classifies and localizes adenoma lesions ($n = 52$) ^a			increased adenoma detection by 23.6%
Wu et al. [13]	Blind spots in esophagogastroduodenoscopy (EGD)	Patients undergoing routine EGD examination at the endoscopy center of Renmin Hospital, Wuhan University (Aug. 2018–Oct. 2018)	<i>WISENSE</i> deep neural network and reinforcement learning system ($n = 153$) The AI system provides real-time feedback on blind spots, scoring of endoscopy based on the number of observed sites, and photo documentation of each inspected site during EGD	Routine EGD without AI <i>WISENSE</i> system ($n = 150$)	Primary outcome: blind spot rate in the <i>WISENSE</i> and routine EGD groups	Significantly lower blind spot rate in the <i>WISENSE</i> arm, compared to the control arm (5.86% vs 22.45%, $p < 0.001$) Significantly longer mean inspection rate in <i>WISENSE</i> arm vs control (5.03 mins vs 4.24 mins, $p < 0.001$). No significant difference observed in the completeness of documentation (72% vs 79%, $p = 0.11$) <i>WISENSE</i> significantly reduced the number of patients being ignored in the lesser curvature of the middle-upper body, as compared to the control (risk difference, -47.5%, $p < 0.001$)
Kaura et al. [14]	Paroxysmal atrial fibrillation (PAF)	Patients aged ≥ 18 years diagnosed with ischemic	iRhythm Technologies <i>Zio</i> ® patch ($n = 56$)	Conventional short-duration Holter ECG monitoring only, at a	Detection of PAF lasting at least 30 s within 90 days for	Significantly greater rate of PAF detection with <i>Zio</i> ® patch, as

<p>non-lacunar stroke or TIA within the past 72 h at King's College Hospital NHS Foundation Trust (Feb. 2016–Feb. 2017)</p>	<p><i>Zio® patch</i> is a small wearable device applied to the anterior chest wall for a 14-day interval, providing continuous ECG monitoring for detecting PAF (primary outcome)</p>	<p>physician determined duration, but usually 24 h ($n = 60$)</p>	<p>each study arm Economic modeling of stroke incidence and medical cost savings with <i>Zio® patch</i></p> <p>result in avoiding 10.8 more strokes annually, with associated medical savings of £113,630, increasing to £162,491 over 5 years</p>
<p>Lin et al. [15]</p>	<p>Childhood cataracts</p>	<p>Pediatric patients (aged ≤ 14 years) without definitive cataract diagnosis or history of previous eye surgery, at five ophthalmic clinics in China (Aug. 2017–May 2018)</p>	<p>Consultation, with childhood cataract AI platform, CC-Cruiser ($n = 175$)</p> <p>Slit lamp ocular photographs uploaded onto CC-Cruiser website, generating a diagnosis (normal vs cataracts), comprehensive metrics (opacity area, density, location), and treatment recommendation (surgery vs follow-up)</p> <p>Routine senior consultant review of patient in clinic ($n = 175$)</p> <p>The diagnostic performance of CC-Cruiser. As compared to expert consultants</p> <p>Disease severity and treatment recommendations, time required to make a diagnosis and patient satisfaction</p> <p>compared to the expert consultant control group (99% and 97%, respectively). Odds ratio 0.06, $p < 0.001$</p> <p>CC-Cruiser had a significantly reduced mean time for diagnosis compared to the control arm.</p> <p>Mean difference of 5.74 min ($p < 0.001$). Patients reported satisfaction with medical service quality and time-saving ability</p>

(continued)

Table 1 (continued)

Publication	Disease Context	Population	Intervention	Control	Outcome	Summary of results
Disease prediction						
Wijnberge et al. [16]	Intraoperative hypotension	Patients scheduled for elective noncardiac surgery at the Amsterdam University Medical Centers, Location AMC, Netherlands (May 2018–Mar. 2019)	Machine learning-derived early warning system ($n = 31$ analyzed) The system collates 23 metrics from the arterial pressure waveform, generating a numerical percentage risk of intraoperative hypotension within the next 15 min	Standard intraoperative care and monitoring by the anesthesics team ($n = 29$)	Time-weighted average of hypotension during surgery	Significantly reduced median time-weighted average ($p = 0.001$) and median time per patient ($p < 0.001$) of intraoperative hypotension in the machine learning system group, as compared to the control arm
Jaroszewski et al. [17]	Risk assessment and crisis intervention	Users of the <i>Koko</i> app [18] (signed up between Aug. 2017 and Sep. 2017). <i>Koko</i> includes an ML component which interprets semantic information and classifies user posts as “crisis” or “not crisis.” Those with posts classed as “crisis” were included in this study	An App (<i>Koko</i>) which provides access to an anonymous peer-to-peer mental health support community. Within the App, an ML classifier provides automated risk assessment and automatically directs those classified as “in crisis” to an intervention platform (barrier reduction intervention (BRI)) which aims to increase the use of crisis resources and support by lowering barriers to intervention access ($n = 775$)	Standard access to <i>Koko</i> mental health app, without barrier reduction intervention (BRI) ($n = 805$)	Likelihood of the user to use crisis resources	Among the participants providing follow-up data, participants receiving BRI were 23% more likely to use crisis services, as compared to control arm (48.9% vs 39.8%, $p = 0.02$)

Therapeutic treatment					
Nimri et al. [19]	Insulin dose optimization	Patients with type 1 diabetes mellitus (enrolled between Nov. 2017 and Jul. 2019)	AI-DSS; artificial intelligence-based decision support system ($n = 54$) The AI system provides automated insulin therapy adjustment recommendations every 3 weeks, based on continuous glucose monitoring	Physician-guided glucose-level control and insulin dose adjustment ($n = 54$)	The percentage of time spent within the target glucose range
Labovitz et al. [20]	Adherence to anticoagulation therapy	Patients diagnosed with ischemic stroke and on anticoagulation therapy. 12-week study (enrolled between Mar. 2015 and Apr. 2016)	Adherence to anticoagulation therapy using the <i>AiCure</i> app ($n = 15$) The app applies facial recognition to confirm the patient's identity, identifies the medication, and confirms ingestion. Clinic staff are notified if doses were missed, late, or used incorrectly	No daily monitoring of anticoagulation therapy ($n = 12$ analyzed)	Glycemic control (measured through time spent within the target glucose range) with AI-DSS was reported as statistically non-inferior to the physician control arm ($50.2 \pm 11.1\%$ vs $51.6 \pm 11.3\%$, $p < 0.001$) Based on plasma drug concentrations, adherence reported as 100% (15/15) with <i>AiCure</i> and 50% (6/12) in the control arm

R-FCN region-based fully connected convolutional neural network, *CADe* automatic computer-aided polyp detection (*CADe*) system (EndoScreener, Vision AI, Shanghai, China), *AQCS* automatic quality control system, *ENDOANGEL* real-time quality improvement system for monitoring withdrawal speed and timings for intubation and withdrawal in colonoscopy, *ADR* adenoma detection rate, *PAF* paroxysmal atrial fibrillation, *BRI* barrier reduction intervention, *AI-DSS* artificial intelligence-based decision support system, *EGD* esophagogastroduodenoscopy, *WISENSE* real-time quality improvement system for blind spots, *EGD* timing, and photo documentation.

^aLui et al. (2021) (nonrandomized trial): The same 52 patients received both control and intervention endoscopy. First, routine colonoscopy was performed and then during the withdrawal phase the AI system was activated

The CADe deep learning system was connected to the endoscopy processor, providing real-time feedback through alert boxes and sound alarms upon polyp detection by the AI system. In the control group, a sham system simulated alerts on polyp-like non-polyp structures (including wrinkled mucosa, undigested debris, feces, and bubbles), masking operating endoscopists to the allocated study arm. The trial concluded a significantly greater ADR in the CADe arm, relative to the control arm (odds ratio 1.36, $p = 0.030$).

Su et al. evaluated the effectiveness of an automatic quality control system (AQCS) which supervises the withdrawal phase of endoscopy once the scope reached the level of the caecum [10]. The system provided additional information on the monitor viewed by the endoscopist, including a real-time timer display, audio alerts to encourage slower withdrawal speeds, audio alerts for mucosal cleaning/liquid suction at inadequately exposed bowel segments, and alert boxes for detected polyps. In addition to reporting statistically superior ADRs (0.289 vs 0.165, $p < 0.001$), this study found the withdrawal time was significantly increased with the AI system, relative to the control arm (7.03 min vs 5.68 min, $p < 0.001$). The adequate bowel preparation rate was also reported as significantly superior in the intervention arm (87.34% vs 80.63%, $p = 0.023$). Previous literature has suggested that higher ADRs can be achieved with longer inspection times during the withdrawal phase, most likely by increasing frame quality and encouraging comprehensive assessment [23, 24].

Gong et al. evaluated an AI detection system (ENDOANGEL), which provides real-time feedback on withdrawal speeds and times to the operator, delivering prompts to encourage care during the withdrawal phase to reduce blind spots [11]. Its efficacy was investigated in a single-blinded randomized controlled study ($n = 704$). This study found that ENDOANGEL significantly improved ADRs with the AI system (odds ratio 2.30, $p = 0.001$), in comparison to unassisted routine colonoscopy.

In a smaller, more recent, nonrandomized study of 52 participants receiving AI-assisted colonoscopy, Lui et al. showed the ADR could

be increased by 23.6% with the AI system's assistance and was able to detect at least 1 missed adenoma in 14 patients [8].

An AI System for Detecting Blind Spots During Esophagogastroduodenoscopy

Esophagogastroduodenoscopy (EGD) is a key investigation for the diagnosis of lesions of the upper gastrointestinal tract. However, the discovery of gastric precursor lesions and cancers ultimately depends upon the quality of EGD and the insight of the performing endoscopist.

Given the variability of its performance, a real-time quality system for monitoring blind spots during EGD has been developed. WISENSE, a real-time detection system based on deep convolutional neural networks and reinforcement learning, is designed to detect anatomical locations seen during EGD (i.e., esophagus, antrum, duodenal bulb, etc.) and provide a rating of adequate coverage (good, excellent, and perfect when certain percentage of sites were visualized). This feedback is intended to help users be aware of blind spots during EGD [13]. In an RCT, patients randomized to WISENSE-assisted EGD ($n = 153$) received real-time AI feedback during the procedure: monitoring of blind spots on a virtual stomach model, timing, and percentage scoring based on the number of observed sites and blind spots, in addition to photo documentation of each inspected site. Routine non-assisted EGD was performed in patients randomized to the control group ($n = 150$).

WISENSE showed the ability to improve quality of EGD in the trial, monitoring blind spots with an average accuracy of 90.4% and providing a significantly lower blind spot rate (5.86%) compared to the control group (22.46%) ($p < 0.001$).

An AI System for Detecting Paroxysmal Atrial Fibrillation

Atrial fibrillation (AF) is recognized to increase cardiovascular risk, including the risk of ischemic stroke and transient ischemic attack (TIA) [25]. However, AF can often be paroxysmal (paroxysmal atrial fibrillation (PAF)) and may be missed by single or 24-h Holter ECG monitoring [26]. This can lead to idiopathic reporting of

stroke/TIA events, delaying prophylactic anti-coagulation therapy for AF.

A 14-day ECG monitoring patch (Zio[®] Patch, iRhythm Technologies) was developed that can be applied and kept in situ on the anterior chest wall for up to 14 consecutive days [27]. The ECG trace is then interpreted by a deep learning model which detects episodes of AF. AF detection rates by the Zio patch were compared to conventional Holter monitoring in an open-label randomized trial (Zio patch $n = 56$, Holter monitoring $n = 60$) [14]. AF detection was found to be significantly higher with the Zio patch compared to Holter monitoring (odds ratio 8.9, $p = 0.026$). Additionally, the authors carried out cost-effectiveness analysis which projected savings in direct medical and outpatient follow-up appointment costs amounting to annual medical cost savings of £113,630, increasing to £162,491 over a 5-year period.

An AI System for Diagnosing Childhood Cataracts

In the absence of early diagnosis and treatment, childhood cataracts can lead to irreversible vision loss [28, 29]. CC-Cruiser, an ophthalmic AI platform, was developed to automate diagnosis and treatment decision-making for this disease [15].

The accuracy of cataract diagnosis and treatment recommendations made by CC-Cruiser were investigated in a multicenter randomized controlled trial [15]. Slit lamp ocular photographs of patients randomized to the intervention group ($n = 175$) were uploaded onto CC-Cruiser. Upon image processing by the AI system, output parameters were reported: the diagnosis (normal lens versus cataract), statistical measures for disease severity (lens opacity, density, and location), and suggestions for treatment (follow-up or surgical). Patients randomized to the control group ($n = 175$) were reviewed in regular ophthalmic clinics by senior consultants. After the receipt of an initial diagnosis, all patients received a gold standard diagnosis by an expert panel of three cataract specialists. The study found a significantly inferior diagnostic accuracy and treatment recommendation performance by CC-Cruiser (87.4% and 70.8%, respectively), in comparison to the senior consultants (99.1% and 96.7%,

respectively), $p < 0.001$ for both. However, a significantly reduced mean time for diagnosis was reported with CC-Cruiser, with a mean difference of 5.74 min ($p < 0.001$), relative to the senior consultant arm. Patients also reported satisfaction with its medical service quality and time-saving ability.

AI Systems for Disease Prediction

An AI Early Warning System for Detecting Intraoperative Hypotension

Intraoperative hypotension during noncardiac surgery has been associated with an increased risk of postoperative complications including myocardial injury, acute kidney injury, and increased mortality [30–32]. To optimize intraoperative hemodynamics, Wijnberge et al. evaluated the effectiveness of an AI early warning system, which analyses arterial pressure waveform from arterial catheter pressure readings to derive a percentage risk of hypotension within the next 15 min [16]. A risk score in excess of 85% activates a sound alarm and flickering light, alerting the anesthesiologist and encouraging corrective action. In this single-center randomized clinical trial (HYPE trial [16]), patients were randomized to the intervention group ($n = 31$), where the ML system performed alerts, or the control group ($n = 29$), where standard care was provided through monitoring by the anesthesiologist. The study reported a lower median average time of hypotension for patients randomized to the machine learning group (8.0 min) compared to the control group (32.7 min), amounting to a statistically significant median difference ($p < 0.01$).

An AI System for Mental Health Risk Assessment

Natural language processing (NLP) is a subfield of AI and ML which aims to computationally read, understand, and derive meaning from text. It is the technique upon which “chatbots” are built, including those designed to conduct health consultations or understand clinical vignettes [33, 34].

This is the technique underlying the mental health app Koko [18]. Koko provides an anonymous peer-to-peer online social network

community where users can create messages. Among its services includes an AI algorithm that interprets the semantic content of user posts on the app and assigns codes to classify the users' risk (as "in crisis" or "not in crisis"). The intervention being assessed in this RCT was a barrier reduction intervention (BRI) which was provided to users detected as "in crisis." [17] The BRI involves presenting a list of crisis resources to users and asking whether they would like to use the resources shared. It continues to explore the users' reasons for not wanting to access intervention. The intervention arm in this RCT included 775 participants, and the control arm (where the app interaction terminates at the point where crisis resources are presented as a list) included 805 participants.

Among the 652 participants providing follow-up data (control arm $n = 327$, intervention arm $n = 325$), a 23% increase ($p = 0.02$) in the use of crisis services in the intervention arm was reported.

AI Systems for Adjusting Therapeutic Treatment

An AI System for Optimizing Insulin Dose

Achieving optimal glycemic control is essential for reducing diabetes-related complications. However, even with insulin pumps and continuous glucose monitoring (CGM) technologies, achieving glycemic goals can prove challenging for patients with insulin-dependent diabetes [35].

An artificial intelligence-based decision support system (AI-DSS) provides health professionals with insulin therapy adjustment recommendations for type 1 diabetic patients who are using CGM and insulin pumps. The effectiveness of the AI-DSS for achieving glycemic control was studied in a multicenter randomized trial where participants were randomized to receive remote insulin dose adjustment every 3 weeks by either AI-DSS ($n = 54$) or a physician ($n = 54$) [19]. The primary outcome was the percentage of time spent within the target glucose range. Glycemic control with AI-DSS was reported as statistically non-inferior to the physician control arm ($50.2 \pm 11.1\%$ vs $51.6 \pm 11.3\%$,

$p < 0.0000001$), studying the times spent within the target glucose range ($70\text{--}180 \text{ mg dl}^{-1}$).

An AI System for Monitoring Drug Adherence

Direct oral anticoagulants (DOACs) offer a favorable risk-benefit profile for anticoagulation therapy, without the need for routine monitoring given its specificity, wide therapeutic window, minimal interactions, and fixed dosing regimen [36]. However, this in turn intensifies the importance of patient compliance and adherence to their prescription. Given that directly observed therapy has demonstrated benefit for optimizing treatment adherence [37, 38], the role of automated systems for monitoring adherence to anticoagulation therapy in stroke patients has been considered.

AiCure, an AI-based app, utilizes software algorithms to identify the patient and their medication and provides visual confirmation of its ingestion [20]. The patient-sensitive information is encrypted in real time, with patient users and clinical staff being alerted of incorrect medication compliance (late, missed, or incorrect dosing).

The effectiveness of AiCure for optimizing adherence to anticoagulation therapy was prospectively studied in a small, single-site randomized study [20]. Patients with recently diagnosed ischemic stroke were randomized to either the intervention group (AI monitoring, $n = 15$) or a control group (no daily monitoring, $n = 13$). One patient from the control group withdrew participation from the study and was not included in the analysis. The trial demonstrated improved adherence to anticoagulation therapy with AiCure, with plasma drug concentrations indicating 100% adherence with the AI system, by comparison to only 50% adherence in the control group.

Reporting Standards for AI Clinical Trials

The critical appraisal of clinical trials is an essential part of evidence-based practice. Reviewers can assess the quality, value, and relevance of a clinical trial by considering the way it was designed, conducted, and analyzed, to formulate

a judgment on the internal and external validity of its findings. This process supports relevant stakeholders when making considered decisions about whether an intervention should be approved and commissioned.

To help ensure accurate, complete, and transparent reporting of health research, reporting guidelines are a helpful aid for establishing minimum reporting standards. For clinical trials, the SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) [39] 2013 and CONSORT (Consolidated Standards of Reporting Trials) [40] statements have ensured minimum reporting guidelines for RCTs. These guidelines have been widely endorsed by the International Committee of Medical Journal Editors (ICMJE) [41] and medical journals, which require authors to comply with them at the point of submission [42].

The SPIRIT 2013 [39] and CONSORT 2010 [40] statements were designed specifically for clinical trials evaluating all health interventions, without limitations on the type of intervention. Over time, a number of extensions have been proposed to cover specific trial designs and interventions, including CONSORT-PRO [43], CONSORT Extension for Cluster Trials [44], CONSORT Extension for Pilot and Feasibility Trials [45], CONSORT Extension for Pragmatic Trials [46], SPIRIT Extension for trials in Child Health: SPIRIT-C [47], and SPIRIT-Path [48].

New Reporting Guidelines for AI Trials to Reflect the New Epoch

Two recently published systematic reviews of studies evaluating AI models for healthcare have highlighted major gaps in the reporting of these studies [4, 5]. The risk that AI interventions could be approved and commissioned for use based on incomplete information highlights the need for new, AI-specific reporting guidance. To address this need, the SPIRIT-AI and CONSORT-AI Steering Group announced in October 2019 an initiative to develop evidence-based, international consensus-based, AI-specific minimum reporting guidelines for AI clinical trials [49].

These guidelines are an extension of the existing SPIRIT 2013 and CONSORT 2010 guidelines and were developed using the EQUATOR (Enhancing the Quality and Transparency of Health Research) network framework [50], using Delphi methodology and an international multidisciplinary consortium, to provide the first international standards for clinical trials of AI health interventions. The SPIRIT-AI [51] and CONSORT-AI [52] extension guidelines include 14 and 15 new items considered sufficiently important for AI interventions for clinical trial protocols, respectively, that should be reported in addition to the core SPIRIT 2013 [39] and CONSORT 2010 [40] items.

Key recommendations include asking investigators to provide clear descriptions of the AI intervention enough to allow replication, including the inclusion and exclusion criteria at the level of the input data and participants, instructions and skills required for use, the clinical setting in which the AI intervention is integrated and its intended users (such as healthcare professionals, patients, and the public), the handling of inputs (such as images) and outputs (such as a probability of disease or classification) of the AI intervention, the expected human-AI interaction, and providing analyses of errors. Authors need to specify the version of the AI system used and state whether this changed during the trial or differs from previous validations; describe how images were acquired, selected, and preprocessed before analysis by the AI system; and report the eligibility criteria at both the level of participants and at the level of the input data.

Alongside SPIRIT-AI and CONSORT-AI, several other AI-specific reporting guidelines are in development in 2021. The TRIPOD-AI [53] (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis—Artificial Intelligence) and STARD-AI [54] (Standards for Reporting Diagnostic accuracy studies—Artificial Intelligence) are under development. The TRIPOD-AI guidelines aim to set reporting standards for learning prediction algorithms and harmonize terminology [53], while the STARD-AI extension guidelines aim to improve the reporting of AI-specific diagnostic accuracy studies [54].

Editorial teams of medical journals and peer reviewers are instrumental in ensuring that published studies are reported with the utmost transparency and completeness to promote well-informed readerships. The extent to which journals adopt, endorse, and require authors to comply with these standards largely determines their impact, as observed with the SPIRIT 2013 and CONSORT 2010 statements. The extension guidelines will assist editors, peer-reviewers, and journal readers to understand, interpret, and critically appraise the quality of the AI trial design and potential for bias in reported outcomes.

Future Challenges

AI reporting guidelines will aid in propelling the field forward and will therefore evolve to keep pace with AI's rapid evolution. The recommendations are most relevant to the state of AI in 2020 (which consisted mainly of disease diagnostic algorithms). Advances in computational techniques, including "adaptive/updating algorithms," which continue to "learn" as they are updated or tuned on new training data, will bring new opportunities for innovation that benefits patients (► Chap. 29, "Meta Learning and the AI Learning Process"). However, they may be accompanied by new challenges around study design and reporting, which will need to be addressed by updating the guidelines to continue to ensure transparency, minimize potential biases, and enhance the trustworthiness and generalizability of results.

References

- State of AI Report 2020. <https://www.stateof.ai/>. 2020. Accessed 14 Feb 2021.
- Accenture. Artificial intelligence: healthcare's new nervous system. 2017. https://www.accenture.com/_acnmedia/PDF-49/Accenture-Health-Artificial-Intelligence.pdf#zoom=50. Accessed 12 Feb 2021.
- Muehlematter UJ, Daniore P, Vokinger KN. Health policy approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis; n.d. [https://doi.org/10.1016/S2589-7500\(20\)30292-2](https://doi.org/10.1016/S2589-7500(20)30292-2).
- Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health*. 2019;1:e271–97. [https://doi.org/10.1016/S2589-7500\(19\)30123-2](https://doi.org/10.1016/S2589-7500(19)30123-2).
- Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies in medical imaging. *BMJ*. 2020;368 <https://doi.org/10.1136/bmj.m689>.
- Sibbald B, Roland M. Understanding controlled trials: why are randomised controlled trials important? *BMJ*. 1998;316:201. <https://doi.org/10.1136/bmj.316.7126.201>.
- Perry Wilson F, Martin M, Yamamoto Y, Partridge C, Moreira E, Arora T, et al. Electronic health record alerts for acute kidney injury: multicenter, randomized clinical trial. *BMJ*. 2021;372 <https://doi.org/10.1136/bmj.m4786>.
- Lui TKL, Hui CKY, Tsui VWM, Cheung KS, Ko MKL, Foo DCC, et al. New insights on missed colonic lesions during colonoscopy through artificial intelligence-assisted real-time detection (with video). *Gastrointest Endosc*. 2021;93:193–200.e1. <https://doi.org/10.1016/j.gie.2020.04.066>.
- Wang P, Liu X, Berzin TM, Glissen Brown JR, Liu P, Zhou C, et al. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADAe-DB trial): a double-blind randomised study. *Lancet Gastroenterol Hepatol*. 2020;5:343–51. [https://doi.org/10.1016/S2468-1253\(19\)30411-X](https://doi.org/10.1016/S2468-1253(19)30411-X).
- Su JR, Li Z, Shao XJ, Ji CR, Ji R, Zhou RC, et al. Impact of a real-time automatic quality control system on colorectal polyp and adenoma detection: a prospective randomized controlled study (with videos). *Gastrointest Endosc*. 2020;91:415–424.e4. <https://doi.org/10.1016/j.gie.2019.08.026>.
- Gong D, Wu L, Zhang J, Mu G, Shen L, Liu J, et al. Detection of colorectal adenomas with a real-time computer-aided system (ENDOANGEL): a randomised controlled study. *Lancet Gastroenterol Hepatol*. 2020;5:352–61. [https://doi.org/10.1016/S2468-1253\(19\)30413-3](https://doi.org/10.1016/S2468-1253(19)30413-3).
- Wang P, Berzin TM, Glissen Brown JR, Bharadwaj S, Becq A, Xiao X, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut*. 2019;68:1813–9. <https://doi.org/10.1136/gutjnl-2018-317500>.
- Wu L, Zhang J, Zhou W, An P, Shen L, Liu J, et al. Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy. *Gut*. 2019;68:2161–9. <https://doi.org/10.1136/gutjnl-2018-317366>.
- Kaura A, Sztriha L, Chan FK, Aeron-Thomas J, Gall N, Piechowski-Jozwiak B, et al. Early prolonged

- ambulatory cardiac monitoring in stroke (EPACS): an open-label randomised controlled trial. *Eur J Med Res.* 2019;24. <https://doi.org/10.1186/s40001-019-0383-8>.
15. Lin H, Li R, Liu Z, Chen J, Yang Y, Chen H, et al. Diagnostic efficacy and therapeutic decision-making capacity of an artificial intelligence platform for childhood cataracts in eye clinics: a multicentre randomized controlled trial. *EClinicalMedicine.* 2019;9:52–9. <https://doi.org/10.1016/j.eclim.2019.03.001>.
 16. Wijnberge M, Geerts BF, Hol L, Lemmers N, Mulder MP, Berge P, et al. Effect of a machine learning-derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the HYPE randomized clinical trial. *JAMA.* 2020;323: 1052–60. <https://doi.org/10.1001/jama.2020.0592>.
 17. Jaroszewski AC, Morris RR, Nock MK. /ine machine learning-driven risk assessment and intervention platform for increasing the use of crisis services. *J Consult Clin Psychol.* 2019;87:370–9. <https://doi.org/10.1037/ccp0000389>.
 18. Koko: About. <https://www.koko.ai/about>. Accessed 10 Feb 2021.
 19. Nimri R, Battelino T, Laffel LM, Slover RH, Schatz D, Weinzimer SA, et al. Insulin dose optimization using an automated artificial intelligence-based decision support system in youths with type 1 diabetes. *Nat Med.* 2020;26:1380–4. <https://doi.org/10.1038/s41591-020-1045-7>.
 20. Labovitz DL, Shafner L, Reyes Gil M, Virmani D, Hanina A. Using artificial intelligence to reduce the risk of nonadherence in patients on anticoagulation therapy. *Stroke.* 2017;48:1416–9. <https://doi.org/10.1161/STROKEAHA.116.016281>.
 21. Bibbins-Domingo K, Grossman DC, Curry SJ, Davidson KW, Epling JW, García FAR, et al. Screening for colorectal cancer: US preventive services task force recommendation statement. *JAMA.* 2016;315:2564–75. <https://doi.org/10.1001/jama.2016.5989>.
 22. Rex DK, Boland CR, Dominitz JA, Giardiello FM, Johnson DA, Kaltenbach T, et al. Colorectal cancer screening: recommendations for physicians and patients from the U.S. Multi-Society Task Force on colorectal cancer. *Am J Gastroenterol.* 2017;112: 1016–30. <https://doi.org/10.1038/ajg.2017.174>.
 23. Lee TJW, Blanks RG, Rees CJ, Wright KC, Nickerson C, Moss SM, et al. Longer mean colonoscopy withdrawal time is associated with increased adenoma detection: evidence from the Bowel cancer screening programme in England. *Endoscopy.* 2013;45:20–6. <https://doi.org/10.1055/s-0032-1325803>.
 24. Barclay RL, Vicari JJ, Doughty AS, Johanson JE, Greenlaw RL. Colonoscopic withdrawal times and adenoma detection during screening colonoscopy. *N Engl J Med.* 2006;355:2533–41. <https://doi.org/10.1056/nejmoa055498>.
 25. Odutayo A, Wong CX, Hsiao AJ, Hopewell S, Altman DG, Emdin CA. Atrial fibrillation and risks of cardiovascular disease, renal disease, and death: systematic review and meta-analysis. *BMJ.* 2016;354: i4482. <https://doi.org/10.1136/bmj.i4482>.
 26. Jabaudon D, Sztajzel J, Sievert K, Landis T, Sztajzel R. Usefulness of ambulatory 7-day ECG monitoring for the detection of atrial fibrillation and flutter after acute stroke and transient ischemic attack. *Stroke.* 2004;35:1647–51. <https://doi.org/10.1161/01.STR.0000131269.69502.d9>.
 27. Zio® by iRhythm UK – Uninterrupted Cardiac Monitoring Service. n.d.. <https://irhythmtech.co.uk/>. Accessed 10 Feb 2021.
 28. Lenhart PD, Courtright P, Wilson ME, Lewallen S, Taylor DS, Ventura MC, et al. Global challenges in the management of congenital cataract: Proceedings of the 4th International Congenital Cataract Symposium held on March 7, 2014, New York, New York. *J AAPOS.* 2015;19:e1–8. <https://doi.org/10.1016/j.jaapos.2015.01.013>. Mosby Inc.
 29. Medsinge A, Nischal KK. Pediatric cataract: challenges and future directions. *Clin Ophthalmol.* 2015;9:77–90. <https://doi.org/10.2147/OPHTHS59009>.
 30. van Waes JAR, van Klei WA, Wijeyesundara DN, van Wolfswinkel L, Lindsay TF, Beattie WS. Association between intraoperative hypotension and myocardial injury after vascular surgery. *Anesthesiology.* 2016;124:35–44. <https://doi.org/10.1097/ALN.0000000000000922>.
 31. Sun LY, Wijeyesundara DN, Tait GA, Beattie WS. Association of intraoperative hypotension with acute kidney injury after elective noncardiac surgery. *Anesthesiology.* 2015;123:515–23. <https://doi.org/10.1097/ALN.0000000000000765>.
 32. Monk TG, Bronsert MR, Henderson WG, Mangione MP, Sum-Ping STJ, Bent DR, et al. Association between intraoperative hypotension and hypertension and 30-day postoperative mortality in noncardiac surgery. *Anesthesiology.* 2015;123:307–19. <https://doi.org/10.1097/ALN.0000000000000756>.
 33. Richens JG, Lee CM, Johri S. Improving the accuracy of medical diagnosis with causal machine learning. *Nat Commun.* 2020;11:1–9. <https://doi.org/10.1038/s41467-020-17419-7>.
 34. Razzaki S, Baker A, Perov Y, Middleton K, Baxter J, Mullarkey D, et al. A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis. *ArXiv.* 2018. arXiv:1806.10698v1 [cs.AI].
 35. Miller KM, Foster NC, Beck RW, Bergensta RM, DuBose SN, DiMeglio LA, et al. Current state of type 1 diabetes treatment in the U.S.: updated data from the T1D exchange clinic registry. *Diabetes Care.* 2015;38: 971–8. <https://doi.org/10.2337/dc15-0078>.
 36. Sikorska J, Upchard J. Direct oral anticoagulants: a quick guide. *Eur Cardiol Rev.* 2017; 12:40–5. <https://doi.org/10.1542/ecr.2017.11.2>.
 37. Mirsaiedi M, Farshidpour M, Banks-Tripp D, Hashmi S, Kujoth C, Schraufnagel D. Video directly observed therapy for treatment of tuberculosis is

- patient-oriented and cost-effective. *Eur Respir J.* 2015;46:871–4. <https://doi.org/10.1183/09031936.00011015>.
38. Hart JE, Jeon CY, Ivers LC, Behforouz HL, Caldas A, Drobac PC, et al. Effect of directly observed therapy for highly active antiretroviral therapy on virologic, immunologic, and adherence outcomes: a meta-analysis and systematic review. *J Acquir Immune Defic Syndr.* 2010;54:167–79. <https://doi.org/10.1097/QAI.0b013e3181d9a330>.
39. Chan AW, Tetzlaff JM, Altman DG, Dickensin K, Moher D. SPIRIT 2013: new guidance for content of clinical trial protocols. *Lancet.* 2013;381:91–2. [https://doi.org/10.1016/S0140-6736\(12\)62160-6](https://doi.org/10.1016/S0140-6736(12)62160-6).
40. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *Int J Surg.* 2011;9:672–7. <https://doi.org/10.1016/j.ijsu.2011.09.004>.
41. International Committee of Medical Journal Editors. Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals: updated December 2019. <http://www.icmje.org/icmje-recommendations.pdf>. Accessed 11 Feb 2021.
42. Moher D, Jones A, Lepage L. Use of the CONSORT statement and quality of reports of randomized trials. *J Am Med Assoc.* 2001;285:1992–5. <https://doi.org/10.1001/jama.285.15.1992>.
43. Calvert M, Blazebey J, Altman DG, Revicki DA, Moher D, Brundage MD. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. *JAMA.* 2013;309:814–22. <https://doi.org/10.1001/jama.2013.879>.
44. Campbell MK, Piaggio G, Elbourne DR, Altman DG. Consort 2010 statement: extension to cluster randomised trials. *BMJ.* 2012;345 <https://doi.org/10.1136/bmj.e5661>.
45. Eldridge SM, Chan CL, Campbell MJ, Bond CM, Hopewell S, Thabane L, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. *BMJ.* 2016;355 <https://doi.org/10.1136/bmj.i5239>.
46. Zwarenstein M, Treweek S, Gagnier JJ, Altman DG, Tunis S, Haynes B, et al. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. *BMJ.* 2008;337:1223–6. <https://doi.org/10.1136/bmj.a2390>.
47. Clyburne-Sherin AVP, Thurairajah P, Kapadia MZ, Sampson M, Chan WWY, Offringa M. Recommendations and evidence for reporting items in pediatric clinical trial protocols and reports: two systematic reviews. *Trials.* 2015;16:417. <https://doi.org/10.1186/s13063-015-0954-0>.
48. Lim SJ. Guidelines for inclusion of pathology-specific assessment and endpoints in clinical trial protocols: the SPIRIT-path extension. *OSF.* 2020; <https://doi.org/10.17605/OSF.IO/E3MF5>.
49. CONSORT-AI and SPIRIT-AI Steering Group. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat Med.* 2019;25:1467–8. <https://doi.org/10.1038/s41591-019-0603-3>.
50. The EQUATOR Network. Reporting guidelines under development. <https://www.equator-network.org/library/reporting-guidelines-under-development/>. Accessed 11 Feb 2021.
51. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ, Ashrafi H, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Lancet Digit Health.* 2020;2:e549–60. [https://doi.org/10.1016/S2589-7500\(20\)30219-3](https://doi.org/10.1016/S2589-7500(20)30219-3).
52. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med.* 2020;26:1364–74. <https://doi.org/10.1038/s41591-020-1034-x>.
53. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet.* 2019;393:1577–9. [https://doi.org/10.1016/S0140-6736\(19\)30037-6](https://doi.org/10.1016/S0140-6736(19)30037-6).
54. Sounderajah V, Ashrafi H, Aggarwal R, de Fauw J, Denniston AK, Greaves F, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: the STARD-AI Steering Group. *Nat Med.* 2020;26:807–8. <https://doi.org/10.1038/s41591-020-0941-1>.



Artificial Intelligence in Medicine: Biochemical 3D Modeling and Drug Discovery

48

Richard Dybowski

Contents

Introduction	661
Predicting the 3D Structures of Proteins	662
Protein Folding	662
Secondary Structure Prediction	663
Ab Initio Tertiary Structure Prediction	663
In Silico Drug Discovery	666
Drug Repurposing	666
Conclusion	669
References	671

Abstract

The shape of a protein is important because it dictates the protein's biological function, and a protein will fold into a configuration that minimizes its overall thermodynamic energy. The total number of possible configurations for a given chain of amino acids grows super-exponentially with the length of the chain; therefore, finding the most stable configuration of a protein is a search over the space of all possible configurations using optimization algorithms with respect to the energies of those configurations. We describe the Deep-Mind's AlphaFold successful approach to this problem. The

second part of the chapter describes a current approach to in silico drug discover based on the use of Generative Adversarial Networks, an important technique in the AI "toolbox."

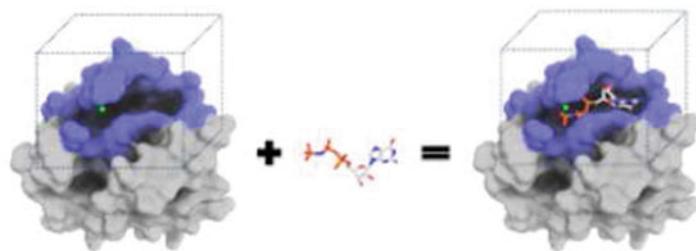
Introduction

In silico drug discovery is the use of computers to find organic molecules that will dock with proteins of interest (Fig. 1). The use of AI for drug discovery is discussed in section "[In Silico Drug Discovery](#)."

The docking of a small molecule with a protein is dependent on the 3D structure of the protein; therefore, in section "[Predicting the 3D Structures of Proteins](#)," we will first look at how computers, and AI, in particular, have been used to predict protein structures.

R. Dybowski (✉)
St John's College, Cambridge, UK
e-mail: rd460@cam.ac.uk

Fig. 1 A docking of guanosine-5'-triphosphate with the H-Ras p21 protein



Predicting the 3D Structures of Proteins

Protein Folding

Proteins are macromolecules consisting of one or more polypeptides. A polypeptide chain will tend to fold into a three-dimensional globular structure (the *native structure*), and the resulting conformation usually has a biological function (Fig. 2). The folding is the result of several forces, and the final structure is determined by the sequence of amino acids forming the polypeptide.

The first step of the overall folding process is the folding of the linear polypeptide chain (the *primary structure*) to a *secondary structure* that consists of α -helices and/or β -sheets. These helices and sheets are a result of intramolecular hydrogen bonds between the amide hydrogen and carbonyl oxygen of the peptide bonds of the primary structure.

The α -helices and β -sheets can contain hydrophilic and hydrophobic portions, and this property of secondary structures aids in the *tertiary structure* of a protein in which the folding occurs so that the hydrophilic sides are facing the aqueous environment surrounding the protein and the hydrophobic sides are facing the hydrophobic core of the protein.

Once the protein's tertiary structure is formed and stabilized by the hydrophobic interactions, there may also be covalent bonding in the form of disulfide bridges formed between two cysteine residues. Tertiary structure of a protein involves a single polypeptide chain; however, additional interactions of folded polypeptide chains give rise to *quaternary structure* formation.

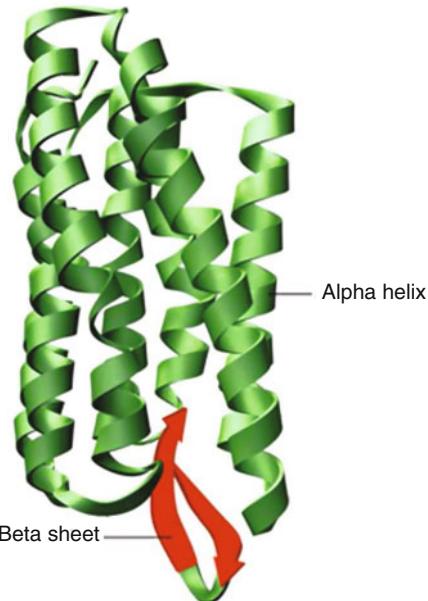


Fig. 2 Bacteriorhodopsin is a protein that acts as a proton pump in bacteria. Its tertiary conformation, which consists of α -helices (green) and β -sheets (red), is essential to its function

Folding is a spontaneous process that is mainly guided by hydrophobic interactions, formation of intramolecular hydrogen bonds, and van der Waals forces, and it is opposed by conformational entropy.

Minimizing the number of hydrophobic side-chains exposed to water is an important driving force behind the folding process [39]. The *hydrophobic effect* is the phenomenon in which the hydrophobic chains of a protein collapse into the core of the protein (away from the hydrophilic environment). In an aqueous environment, the water molecules tend to aggregate around the hydrophobic regions or side chains of the protein,

creating water shells of ordered water molecules [9]. An ordering of water molecules around a hydrophobic region increases order in a system and therefore contributes a negative change in entropy (less entropy in the system). The water molecules are fixed in these water cages, which drives the hydrophobic collapse, or the inward folding of the hydrophobic groups. The hydrophobic collapse introduces entropy back to the system via the breaking of the water cages which frees the ordered water molecules.

Secondary Structure Prediction

One of the earliest algorithms for the prediction of protein secondary structure is the method by Chou and Fasman [8], but this takes into account only the probability that each individual amino acid will appear in a helix, strand, or turn. In contrast, the GOR method [14] takes into account not only the propensities of an individual amino acid to form a particular secondary structure, but also the conditional probability of the amino acid to form a particular secondary structure given that its immediate neighbors have already formed that structure. The GOR method had an accuracy of about only 64%, but one should keep in mind the limited amount of data on protein structures available at the time.

An early application of neural networks to the task of secondary structure prediction was that of Holley and Karplus [21]. This had a maximum overall predictive accuracy of 63%, but the network had only a single hidden layer with two nodes. In contrast, by using an additional hidden layer, Jones [22] achieved 78.3% accuracy, and when deep learning was used via a convolutional neural network, further improvements were made; for example, Wang et al. [55] obtained an accuracy of 84.9%.

β -turns are a particular type of tight turn within a protein consisting of four consecutive residues and they are one of the most common types of nonrepetitive motifs in a protein. β -turns are very significant in protein structure and function partly because, as four-residue reversals, they help in the formation of higher-order structure.

Consequently, their prediction is of importance, and Zhang et al. [59] used a support vector machine to predict β -turns with 77.3% accuracy.

Ab Initio Tertiary Structure Prediction

One way of defining the 3D configuration of a polypeptide based on a sequence S of amino acids is in terms of the relative coordinates \mathbf{x} of all the C carbons present along the polypeptide. Let Ω_x be the space of all possible configurations of \mathbf{x} , then the optimal configuration \mathbf{x}^* of the polypeptide is that for which the resulting energy $G(\mathbf{x})$ is minimal:

$$\mathbf{x}^* = \arg \min_{\mathbf{x} \in \Omega_x} G(\mathbf{x}). \quad (1)$$

This is at the heart of *ab initio* folding prediction; however, the two main problems are the calculation of protein free energy and finding the global minimum of this energy.

Anfinsen's hypothesis [2] is that the 3D structure of a protein in its physiological milieu is the one in which the Gibbs free energy of the whole state is lowest. The Gibbs free energy is given by

$$\Delta G = \Delta H - T\Delta S \quad (2)$$

where enthalpy ΔH is based on bonding energies within proteins including disulfide bonds, hydrogen bonds, van de Waals forces, and electrostatic forces. Entropy ΔS is associated with the hydrophobic effect on proteins, and T is temperature in degrees Kelvin.

A protein structure prediction method must explore the space of possible protein structures which is astronomically large. These problems can be partially bypassed in *homology modeling* [7] and fold recognition methods, in which the search space is pruned by the assumption that the protein in question adopts a structure that is close to the experimentally determined structure of another homologous protein. But what if there is no protein in the Protein Data Bank (PDB) related to the target protein's sequence?

In *fragment assembly* [23], a target protein sequence is deconstructed into small, overlapping fragments. A search of the PDB is performed to identify known structures of similar fragment sequences, which are then assembled into a full-length prediction. The qualities of fragments and their assemblies are assessed by using some form of scoring function that aims to select more native-like protein structures from among the many possible combinations. But note that fragment-assembly techniques do not reliably scale to longer proteins (i.e., ≥ 70 residues).

Predicating the native form of a protein without recourse to fragment assembly or homology modeling has been the grand challenge [11]. *De novo* approaches depend on an effective conformation searching algorithm and good energy functions to build protein tertiary structures, and it is the arrival of deep-learning techniques that have led to significant progress in protein folding prediction [49]. An example of this is the use of a

CNN to predict the torsional angles of a protein from S as a step toward 3D structure prediction [6]. The other steps included residue-contact prediction via pseudo-likelihoods [13] and molecular dynamics simulation [24], but a more eloquent and highly successful approach is arguably that of Deep-Mind's AlphaFold, which we now describe.

AlphaFold

Like most modern prediction algorithms, AlphaFold [47] relies on multiple sequence alignment (MSA). The sequence S of the protein whose structure we intend to predict is compared across a large database. The underlying idea is that if two amino acids are in close contact, mutations in one of them will be closely followed by mutations of the other, in order to preserve protein structure (Figs. 3 and 4).

The central component of AlphaFold is a convolutional neural network that is trained on

Fig. 3 The dihedral torsion angles within residues

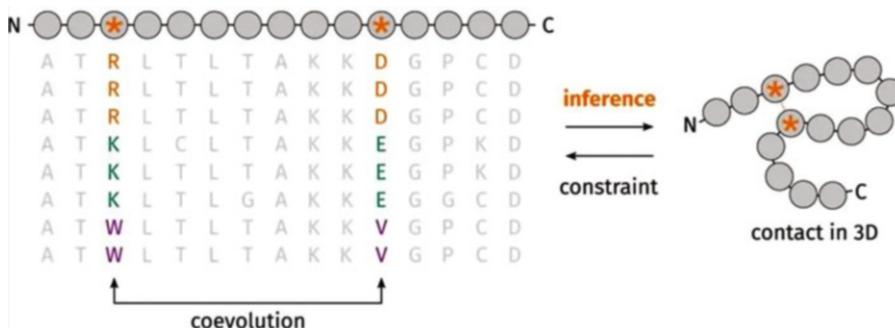
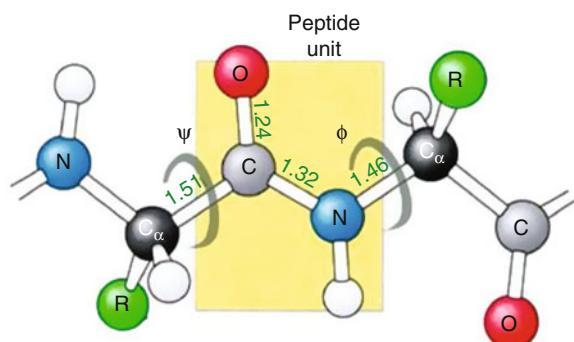


Fig. 4 Protein structure prediction by coevolution. For a given sequence S , homologous sequences can be used to create a multiple sequence alignment MSA (S). Some positions coevolve (orange asterisk) where, for a change at one position, a suitable change at the second position can

be observed. This connection at sequence level implies spatial proximity of both residues. The predicted contacts are used as constraints of a subsequent structure reconstruction in order to find an optimal three-dimensional structure [4, 33]

protein structures to predict the distances $d_{i,j}$ between the C_β atoms of pairs, i, j , of residues of a protein. On the basis of the amino acid sequence S of a protein, and features derived from the MSA (S) of that sequence, the convolutional network predicts a probability distribution $\mathbb{P}(d_{i,j}|S, \text{MAS}(S))$ for every i, j pair. The convolutional network also predicts a probability distribution of backbone torsion angles $\mathbb{P}(\psi, \phi|S, \text{MAS}(S))$.

The backbone atom coordinates \mathbf{x} of a protein (i.e., the relative coordinates of the C_β atoms of the residues) are given as a function $G(\psi, \phi)$ of all the torsion angles.

The total potential $V_{total}(\psi, \phi)$ of a candidate protein structure (ψ, ϕ) with amino acid sequence S is a combination of distance and torsion potentials:

$$V_{total}(\psi, \phi) = V_{distance}(G(\psi, \phi)) + V_{torsion}(\psi, \phi) + V_{score2_smooth}(G(\psi, \phi)),$$

where

$$\begin{aligned} V_{distance}(G(\psi, \phi)) &= - \sum_{i,j, i \neq j} \log \mathbb{P}(d_{i,j}|S, \text{MSA}(S)) \\ &\quad - \log \mathbb{P}(d_{i,j}|\text{length}), \end{aligned} \quad (3)$$

$$V_{torsion}(\psi, \phi) = - \sum_i \log \mathbb{P}(\psi_i, \phi_i|S, \text{MSA}(S)),$$

(ψ_i and ϕ_i are the torsion angles of the i -th residue), and score2_smooth refers to the Rosetta score function [30], which is a linear combination of weighted score terms (listed in Fig. 5) that balance physics-based and statistically derived potentials. This potential is added to prevent steric clashes because it incorporates a van der Waals term E_{vdW} . Term $\mathbb{P}(d_{i,j}|\text{length})$ in Eq. 3 is included to provide a reference distribution that is independent of the protein sequence [48].

As each term of $V_{total}(\psi, \phi)$ is differentiable, it is minimized using gradient descent; however, since the “landscape” of $V_{total}(\psi, \phi)$ over (ψ, ϕ) -space can have many local minima, DeepMind

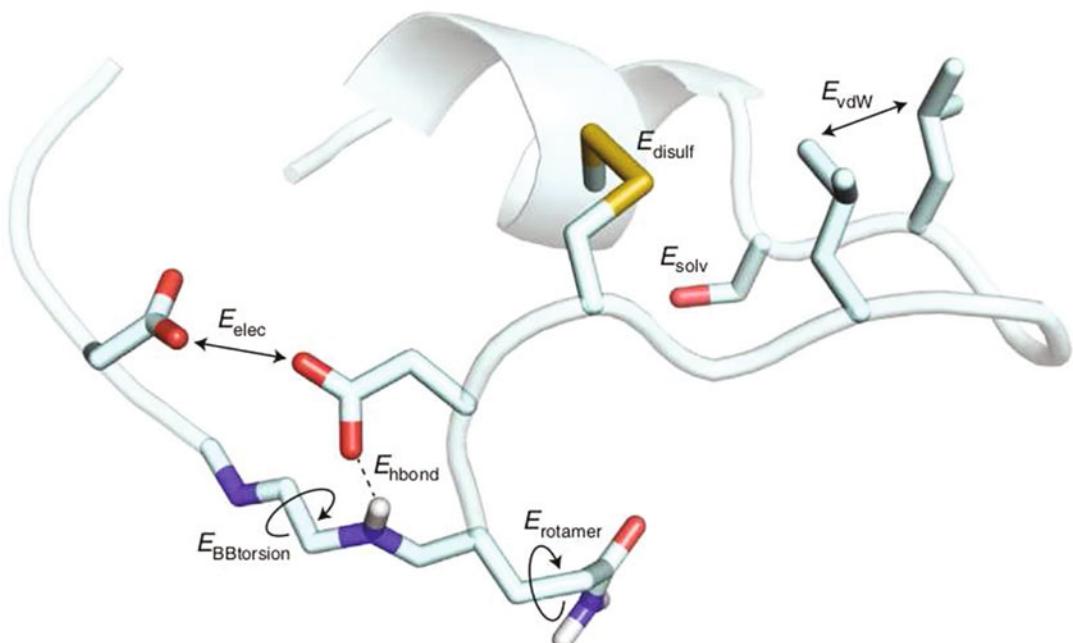


Fig. 5 The elements of Rosetta scoring [30]. E_{vdW} : van der Waals Lennard-Jones potential. E_{hbond} : hydrogen bonding. E_{elec} : electrostatic interaction between charges. E_{disulf} : disulfide bonds between cysteines. E_{solv} : implicit

solvation model. $E_{BBtorsion}$: backbone torsion preferences from main-chain potential. $E_{rotamer}$: side-chain torsion angles from rotamer library. E_{ref} : unfolded state reference energy

restarted the gradient descent using different randomly chosen initial positions in (ψ, ϕ) -space in order to discover the global minimum and thus the optimal protein structure (ψ^*, ϕ^*) for S :

$$V_{\text{total}}(\psi^*, \phi^*) = \arg \min_{x \in \Omega_{(\psi, \phi)}} V_{\text{total}}(\psi, \phi).$$

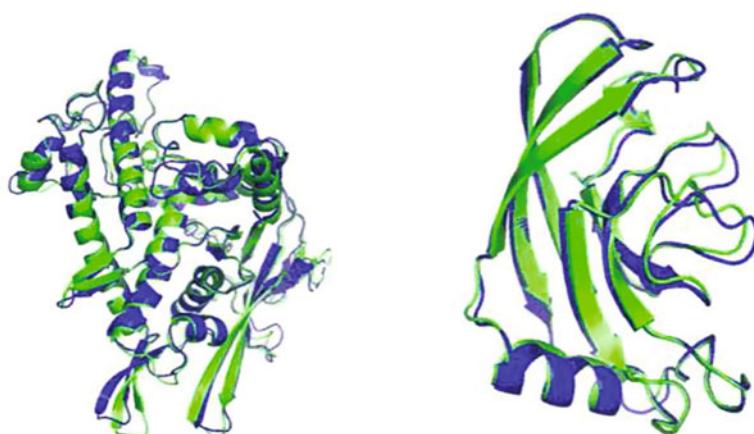
The *Critical Assessment of Structure Prediction* (CASP) is a biennial assessment of protein structure prediction, and it has become the gold standard for assessing predictive techniques for over 20 years. Participants must blindly predict the structure of proteins that have only recently been experimentally determined and have their predictions compared to experimental data.

CASP uses the Global Distance Test (GDT) metric to assess accuracy, which is a percentage ranging from 0 to 100 [58]. AlphaFold achieved a median score of 92.4 GDT across all protein targets, and predictions had a root-mean-square deviation of atomic positions of approximately 1.6 Å, which is comparable to the width of an atom. Figure 6 shows two predictions made by ALPHAFOOLD.

In Silico Drug Discovery

Drug discovery is the process by which new candidate medications are discovered; however, despite advances in pharmacology and a greater understanding of biological systems, drug discovery is still a lengthy, expensive and difficult

Fig. 6 Predictions made by AlphaFold (blue) compared with experimental data (green) for CASP 2020 [10]. The protein on the left is associated with bacterial RNA polymerase (predicted with 90.7 GDT); the one on the right is part of a bacterial adhesion (predicted with 93.3 GDT)



process with low rate of new therapeutic discovery. In 2010, the research and development cost of each new molecular entity was about US\$1.8 billion [40].

Because of the high failure rate of drug development, there is a need for early target validation. This concern has prompted the exploration and implementation of a wide range of computational methods for de novo drug design, and deep-learning neural networks have been shown to be particularly promising in this regard.

Drug Repurposing

Drug repurposing refers to taking a drug that has been developed for one disorder and “repositioning” it to treat another. In contrast to the rising costs, cycle times, and risks associated with drug discovery and development, the repositioning sector has a relatively low company failure rate [37]. In addition, there are 400 million people affected with rare diseases such as cystic fibrosis, but given the nonviability of developing de novo therapies for each of the 8000 rare diseases in existence, drug repositioning offers an alternative approach.

Drug repurposing can now be performed either experimentally or computationally [1], including the use of genome-wide association studies [45].

Deep Neural Networks

Deep neural networks (DNNs) generally outperform other machine-learning methods in

terms of predictive accuracy [15] due to their ability to approximate complex functions by abstracting information hierarchically from data; for example, Ma et al. [31] showed DNNs performing generally better than random forest classifiers across 15 diverse QSAR data sets, including those for CYP P450 3A4 inhibition and human thrombin inhibition. Unterthiner et al. [51] also demonstrated this with regard to drug-target interaction, and this led Vanhaelen et al. [53] to stress the potential of DNNs to drug repurposing. Unterthiner et al. used a multitask DNN to perform QSAR predictions for 5069 targets and 743,336 compounds recorded in the ChEMBL database (Table 1).

ATOMNET [54] is a deep convolutional neural network designed by AtomWise to perform structure-based predictions of protein-ligand activity on target proteins having little, if any, bioactivity data. Wallach et al. found AtomNet to significantly outperform SMINA [29], which is a variant of AutoDock Vina [50]; for example, AtomNet achieved an AUC greater than 0.9 for 57.8% of the targets in the DUDE dataset [36].

Recently, Vamathevan et al. [52] provided a comprehensive description of machine-learning applications to drug discovery, in which they emphasize the results of Mayr et al. [34] that show deep-learning neural networks significantly outperform other methods (Fig. 7) and that the predictive performance of deep learning was in many cases comparable to that of tests performed with *in vitro* assays.

It is important when using machine learning for drug discovery that the ligand-protein data used to test the machine-learning system is not biased [5].

Generative Adversarial Networks

A recent development in the field of neural-based computation is the use of DNNs as data generators, the aim of which is to learn, from a given set of observed data \mathbf{x}_{obs} , a model (the data generator) that emulates a distribution $p_{data}(\mathbf{x})$ from which observed data was sampled.

The usual approaches to drug discovery tend to be restricted to the space Ω_{known} of known chemical compounds, which is a subset of the space of all possible compounds, Ω_{chem} . In contrast, the generative approach allows the discovery of new compounds lying outside of Ω_{known} (Fig. 8).

One type of generator system is the *generative adversarial network* (GAN) [17]. Here, the generator $G(\mathbf{z}; \theta_G)$ begins by generating some data \mathbf{x}_{gen} and a neural network $D(\mathbf{x}_{gen}; \mathbf{x}_{obs}, \theta_D)$ attempts to discriminate between the generated data and the observed data by adjusting its weights θ_D (Fig. 9). Once the neural network succeeds, the generator again generates another set of data to try to challenge the neural network's ability to discriminate between two sets of data (hence "adversarial") by modifying θ_G . This game-theoretic cycle continues via an evaluation function $V(D, G)$ (Fig. 10) until neither the generator nor the neural network can progress any further. More precisely, the training of a GAN involves cycling between minimizing $V(D, G)$ with respect to G and maximizing it with respect to D until the optimal value is obtained, which is at a saddle point. At this point (a Nash equilibrium), the probability mass function of the generator $p(\mathbf{x} | \theta_{gen})$ approximates that from which the observed data was sampled $p_{data}(\mathbf{x})$ [17], and this enables the estimation of data sampled from $p_{data}(\mathbf{x})$.

There are, however, some issues with GANs, including the following:

Table 1 Performance accuracy (in terms of AUC metrics) of a multitask DNN against a support vector machine, binary kernel discrimination, logistic regression, and k -

nearest neighbor. The p -value refers to a paired Wilcoxon test comparing the multitask DNN with the alternative method [51]

Method	AUC	p -value
Multitask DNN	0.830	
Support vector machine	0.816	<0.001
Binary kernel discrimination	0.803	<0.001
Logistic regression	0.796	<0.001
k -Nearest neighbor	0.775	<0.001

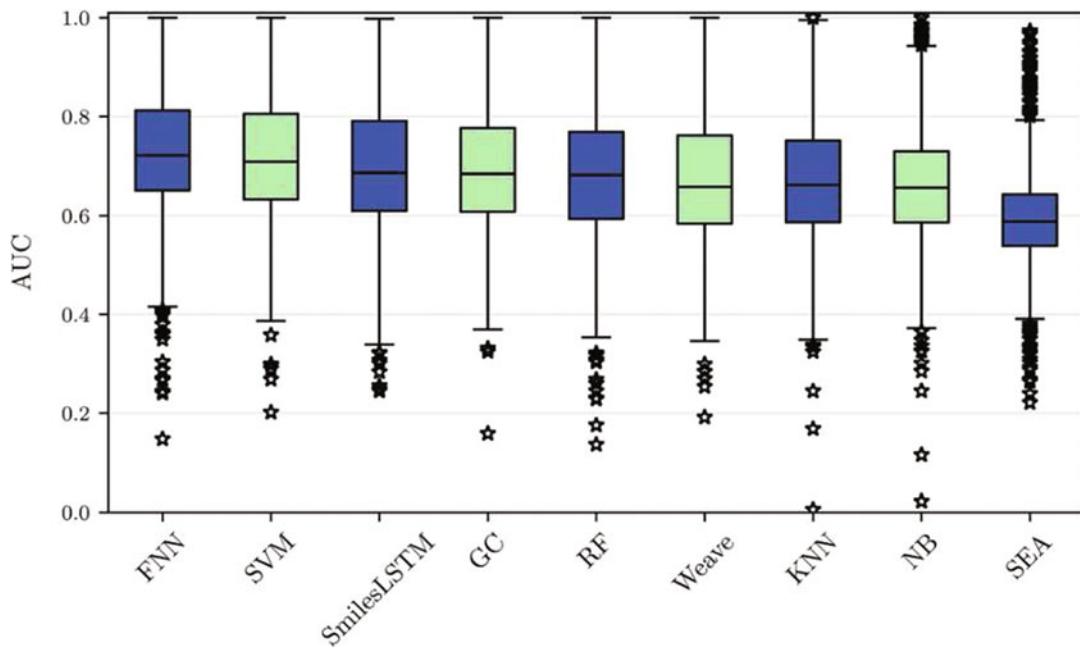


Fig. 7 Performance comparison of drug-target prediction methods [52]. Each method yielded 1310 AUC values from a set of modeled assays. On average, a DNN (FNN) performed best followed by support vector machine (SVM), recurrent neural networks with long short-term

memory (SmilesLSTM), graphical convolution network (GC) [57], random forest (RF), another graphical convolution network (Weave) [26], k-nearest neighbor (KNN), naive bayes (NB), and the similarity ensemble approach (SEA) [27]

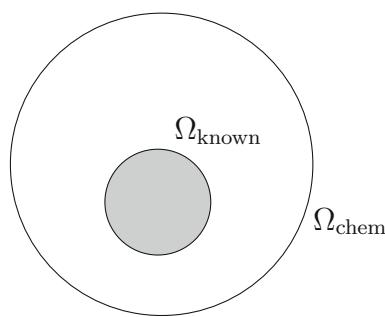


Fig. 8 Ω_{chem} is the space of all possible chemical compounds, and Ω_{known} the set of all known organic compounds. Subset Ω_{known} is not drawn to scale as the cardinality of Ω_{known} is about 60 million whereas as that of Ω_{chem} is about 166 billion [43]

1. There is a need to make sure that neither the generator nor the discriminator becomes too “strong” compared to the other. If the discriminator “wins” and classifies all true data correctly, the error signal will be poor and the generator will not be able to learn from it. If the generator “wins,” it is usually exploiting a nonmeaningful weakness in

the discriminator, which can result in oscillations.

2. A GAN is attempting to move $p_{\text{generator}}(\mathbf{x})$ towards $p_{\text{data}}(\mathbf{x})$, but \mathbf{x} could be very high-dimensional. In this case, it can be better to use a lower-dimensional alternative such as an adversarial autoencoder.

The *adversarial autoencoder* [32] is based on the concept of an autoencoder [3] (Figs. 11 and 12). Unlike the case with GANs, the adversarial autoencoder discriminates with respect to the latent values within the middle layer of an autoencoder, and when Kadurin et al. [25] trained an adversarial autoencoder with 6252 compounds profiled on the MCF-7 cell line (Fig. 13), it generated, and thus predicted, 640 possible new anti-cancer drugs.

In contrast to *variational autoencoders* [28, 42], GANs do not require log-likelihood estimation; however, the cyclic conflict between the adjustment of θ_G for the generator of a GAN and of θ_D for its discriminator can lead to

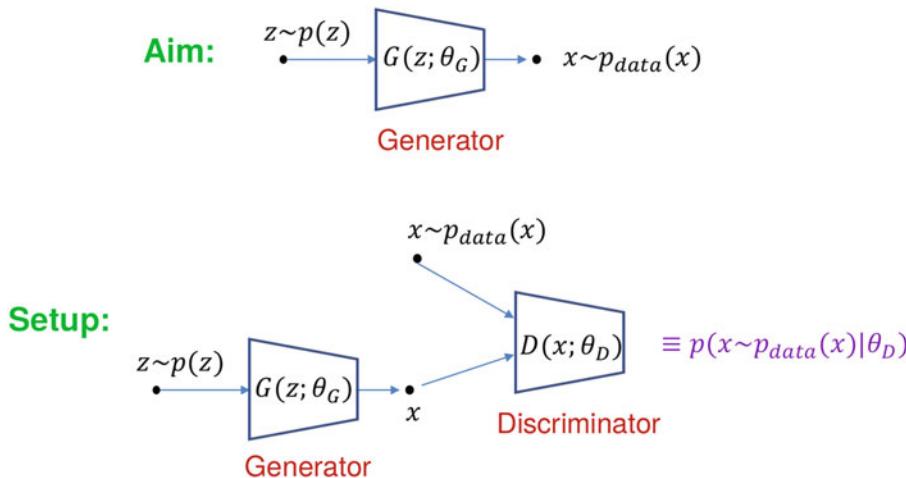


Fig. 9 Schematic of a generative adversarial network

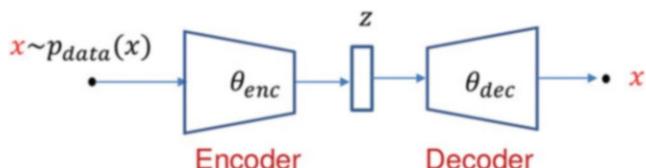
Fig. 10 GAN evaluation function $V(D, G)$

$$V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))]$$

Expected probability that real data will be correctly identified

Expected probability that generated data will be correctly identified

Fig. 11 Schematic of an autoencoder



nonconvergence [35, 44]. In response to this issue, Salimans et al. [44] and Grnarova et al. [18] suggested modifications to GAN structures to mitigate the potential instability of GANs. Olivecrona et al. [38] avoided this issue by using a nonadversarial generative technique based on recurrent neural networks and reinforcement learning. This enabled new molecules to be generated with predefined desirable qualities.

Inputting Molecular Structures

Molecular structures can be mapped bijectively to strings; for example, Smiles [56] is a formal grammar that describes molecules with an alphabet of characters; for example, the Smiles for benzene is c1ccccc1. Given that *recurrent neural networks* [41] are able to learn language models from data

[16], Segler et al. [46] used a recurrent neural network with Long Short-Term Memory [20] to generate the SMILES of molecules that are active against targets of interest (i.e., the receptor 5-HT_{2A}, the parasite *Plasmodium falciparum*, and the bacterium *Staphylococcus aureus*). An alternative approach to using molecular fingerprints such as Smiles is to take the entire 3D structure of a molecule as input [26, 57].

Conclusion

The application of AI to the *ab initio* prediction of protein structures has been critical to the advancement in this field, and the development of AlphaFold is undoubtedly a major scientific

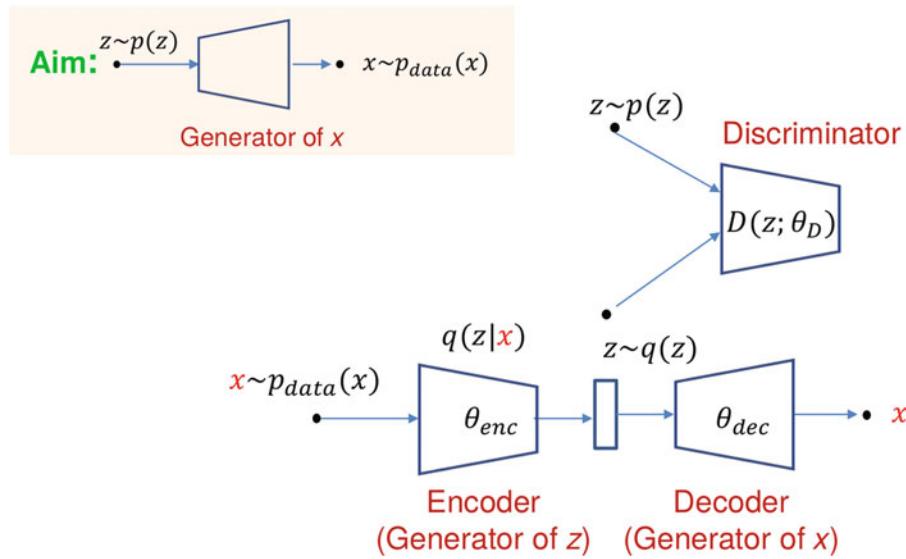


Fig. 12 Adversarial autoencoder

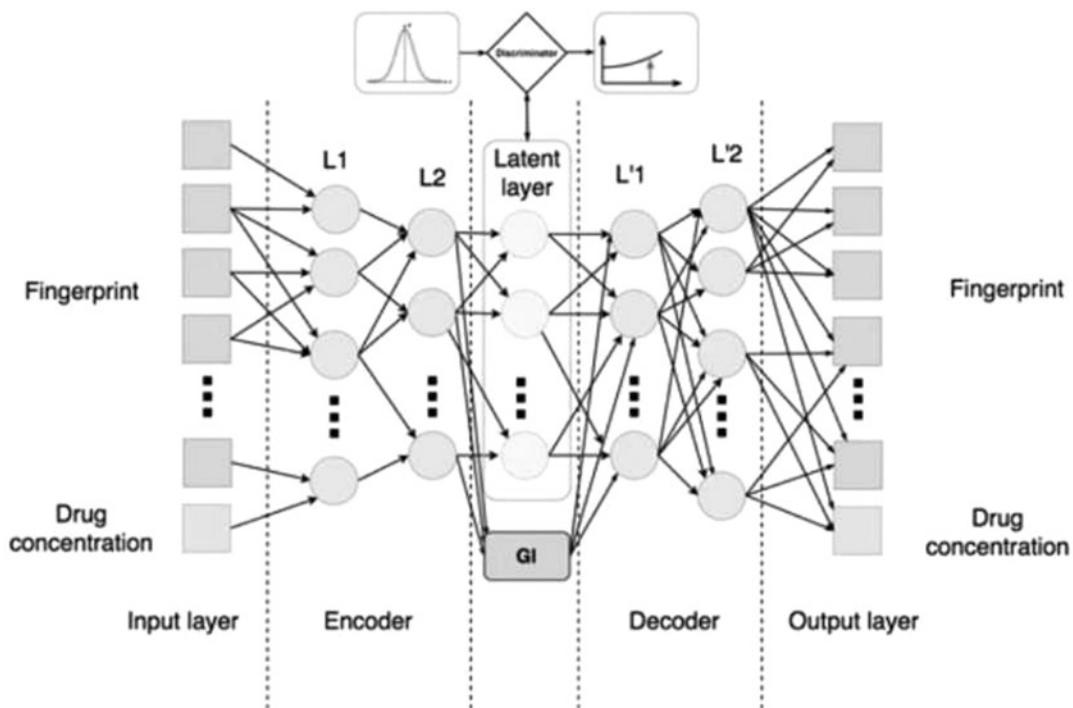


Fig. 13 Adversarial autoencoder used by Kadurin et al. [25]

breakthrough, and it already in use to predict the structures of proteins of SARS-CoV-2, the causative agent of COVID-19 [19].

As for AI and drug discovery, deep neural networks are the leaders. But consider ATOMNET: how can we discover what chemistry ATOMNET

has learnt from the vast amount of ligand-protein-interaction data used to train it for *in silico* drug screening? Being able to answer this question moves drug discovery from prediction to scientific discovery [12].

References

- Alaimo S, Giugno R, Pulvirenti A. Recommendation techniques for drug-target interaction prediction and drug repositioning. In: Carugo O, Eisenhaber F, editors. Data mining techniques for the life sciences. 2nd ed. New York: Springer; 2016.
- Anfinsen C. Principles that govern the folding of protein chains. *Science*. 1973;181(4096):223–30.
- Bengio Y. Neural networks for speech and sequence recognition. London: International Thompson Computer Press; 1996.
- Bittrich S, Schroeder M, Labudde D. StructureDistiller: structural relevance scoring identifies the most informative entries of a contact map. *Sci Rep – Nature*. 2019;9:18517.
- Chen L, Cruz A, Ramsey S, Dickson C, Duca J, Hornak V, ... Kurtzman T. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLoS One*. 2019;14(8):e0220113.
- Cheung N, Yu W. *De novo* protein structure prediction using ultra-fast molecular dynamics simulation. *PLoS One*. 2018;13(11):e0205819.
- Chothia C, Lesk A. The relation between the divergence of sequence and structure in proteins. *EMBO J*. 1986;5:823–6.
- Chou P, Fasman G. Empirical predictions of protein conformation. *Annu Rev Biochem*. 1978;47:251–76.
- Cui D, Ou S, Patel S. Protein-spanning water networks and implications for prediction of protein-protein interactions mediated through hydrophobic effects. *Proteins*. 2014;82(12):33123326.
- DeepMind. AlphaFold: a solution to a 50-year-old grand challenge in biology. 2020. <https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology>. Online: Accessed 2 Dec 2020.
- Dill K, MacCallum J. The protein-folding problem, 50 years on. *Science*. 2012;338:1042–6.
- Dybowski R. Interpretable machine learning as a tool for scientific discovery in chemistry. *New J Chem*. 2020;44:20914–20.
- Ekeberg M, Lökvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins using pseudo-likelihoods to infer Potts models. *Phys Rev E*. 2013;87(1):012707.
- Garnier J, Osguthorpe D, Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol*. 1978;120(1):97–120.
- Goh G, Hodas N, Vishnu A. Deep learning for computational chemistry. *arXiv*, 1701.04503. 2017.
- Goldberg Y. A primer on neural network models for natural language processing. *J Artif Intell Res*. 2016;57:345–420.
- Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, ... Bengio Y. Generative adversarial networks. *arXiv*, 1406.2661. 2014.
- Grnarova P, Levy K, Lucchi A, Hofmann T, Krause A. An online learning approach to generative adversarial networks. *arXiv*, 1706.03269v1. 2017.
- Heo L, Feig M. Modeling of severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2) proteins by machine learning and physics-based refinement. *bioRxiv*. 2020. <https://doi.org/10.1101/2020.03.25.008904>.
- Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9(8):1735–80.
- Holley L, Karplus M. Protein secondary structure prediction with a neural network. *PNAS*. 1989;86(1):152–6.
- Jones D. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*. 1999;292:195–202.
- Jones D, McGuffin L. Assembling novel protein folds from super-secondary structural fragments. *Proteins Suppl*. 2003;6:480–5.
- Jumper J, Freed K, Sosnick T. Maximum-likelihood, self-consistent side chain free energies with applications to protein molecular dynamics. *arXiv*, 161007277. 2016.
- Kadurin A, Aliper A, Kazennov A, Mamoshina P, Vanhaelen Q, Khrabrov K, Zhavoronkov A. The cornucopia of meaningful leads: applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget*. 2017;8(7):10883–90.
- Kearnes S, McCloskey K, Berndl M, Pande V, Riley P. Molecular graph convolutions: moving beyond fingerprints. *J Comput Aided Mol Des*. 2016;30(8):595–608.
- Keiser M, Roth B, Armbruster B, Ernsberger P, Irwin J, Shoichet B. Relating protein pharmacology by ligand chemistry. *Nat Biotechnol*. 2007;25:197–206.
- Kingma D, Welling M. Auto-encoding variational Bayes. *arXiv*, 1312.6114v10. 2013.
- Koes D, Baumgartner M, Camacho C. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J Chem Inf Model*. 2013;53(8):1893–904.
- Leman J, Weitzner B, Lewis S, ... Bonneau R. Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat Methods*. 2020;17:665–80.
- Ma J, Sheridan R, Liaw A, Dahl G, Svetnik V. Deep neural nets as a method for quantitative structure-activity relationships. *J Chem Inf Model*. 2015;55: 263–74.
- Makhzani A, Shlens J, Jaitly N, Goodfellow I, Frey B. Adversarial autoencoder. *arXiv*, 1511.05644. 2016.
- Marks D, Colwell L, Sheridan R, Hopf T, Pagnani A, Zecchina R, Sander C. Protein 3D structure computed

- from evolutionary sequence variation. PLoS One. 2011;6:e28766.
34. Mayr A, Klambauer G, Unterthiner T, Steijaert M, Wegner J, Ceulemans H, ... Hochreiter S. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem Sci*. 2018;9(24):5441–51.
35. Metz L, Poole B, Pfau D, Sohl-Dickstein J. Unrolled generative adversarial networks. arXiv, 1611.02163. 2016.
36. Mysinger M, Carchia M, Irwin J, Shoichet B. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J Med Chem*. 2012;55(14):6582–94.
37. Naylor S, Kauppi D, Schonfled J. Therapeutic drug repurposing, repositioning and rescue. Part II: Business review. *Drug Discov World*. 2015;16:57–72.
38. Olivcrona M, Blaschkey T, Engkvist O, Cheny H. Molecular de-novo design through deep reinforcement learning. arXiv, 1704.07555v1. 2017.
39. Pace C, Shirley B, McNutt M, Gajiwala K. Forces contributing to the conformational stability of proteins. *FASEB J*. 1996;10(1):75–83.
40. Paul S, Mytelka D, Dunwiddie C, Persinger C, Munos B, Lindborg S, Schacht A. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov*. 2010;9(3):203–14.
41. Pearlmutter B. Learning state space trajectories in recurrent neural networks. *Neural Comput*. 1989;1:263–9.
42. Rezende D, Mohamed S, Wierstra D. Stochastic back-propagation and approximate inference in deep generative models. arXiv, 1401.4082v3. 2014.
43. Ruddigkeit L, van Deursen R, Blum L, Reymond J-L. Enumeration of 166 billion organic small molecules in the Chemical Universe Database GDB-17. *J Chem Inform Model*. 2012;52(11):28642875.
44. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training GANs. arXiv, 1606.03498. 2016.
45. Sanseau P, Agarwal P, Barnes M, Pastinen T, Richards J, Cardon L, Mooser V. Use of genome-wide association studies for drug repositioning. *Nat Biotechnol*. 2012;30:317320.
46. Segler M, Kogej T, Tyrchan C, Waller M. Generating focussed molecule libraries for drug discovery with recurrent neural networks. arXiv, 1701.01329. 2017.
47. Senior A, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, ... Hassabis D. Improved protein structure prediction using potentials from deep learning. *Nature*. 2020;577(7792):706–10.
48. Simons K, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol*. 1997;268:209–25.
49. Torrisi M, Pollastri G, Le Q. Deep learning methods in protein structure prediction. *Comput Struct Biotechnol J*. 2020;18:1301–10.
50. Trott O, Olson A. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J Comput Chem*. 2010;31:455–61.
51. Unterthiner T, Mayr A, Klambauer G, Steijaert M, Wegner J, Ceulemans H, Hochreiter S. Deep learning as an opportunity in virtual screening. NIPS Workshop on Deep Learning and Representation Learning. Montreal. 12 December 2014. 2014.
52. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G ... Zhao S. Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov*. 2019;18:463–77.
53. Vanhaelen Q, Mamoshina P, Aliper A, Artemov A, Lezhmina K, Ozerov I, ... Zhavoronkov A. Design of efficient computational workflows for *in silico* drug repurposing. *Drug Discov Today*. 2017;22(2):210–22.
54. Wallach I, Dzamba M, Heifets A. AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. ArXiv, 1510.02855. 2015.
55. Wang S, Peng J, Ma J, Xu J. Protein secondary structure prediction using deep convolutional neural fields. *Sci Rep – Nature*. 2016;6:18962.
56. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Model*. 1988;28(1):31–6.
57. Wu Z, Ramsundar B, Feinberg E, Gomes J, Geniesse C, Pappu A, ... Pande V. MoleculeNet: a benchmark for molecular machine learning. *Chem Sci*. 2018;9:513–30.
58. Zemla A. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res*. 2003;31(13):33703374.
59. Zhang Q, Yoon S, Welsh W. Improved method for predicting β -turn using support vector machine. *Bioinformatics*. 2005;21(10):2370–4.



Namki Hong, Yurang Park, Seng Chan You, and Yumie Rhee

Contents

Introduction	674
AI Applications in Diabetes Mellitus	675
AI-Driven Diabetes Care: Changing the Landscape	675
Prediction of Diabetes Risk	676
Retinopathy Detection	677
Prediction of Diabetic Complications	677
Continuous Glucose Monitoring and Closed-Loop Artificial Pancreas System	678
Therapeutic Lifestyle Modification	678
AI Applications in Bone and Mineral Disorders	679
Fracture Identification	679
Opportunistic Screening of Osteoporosis and Sarcopenia	679
Fracture Risk Assessment	680
Finding Novel Biomarkers Related to Bone Metabolism	680
AI Applications in Thyroid Disorders	681
AI Application in Thyroid Cancer	681
AI Application in Functional Thyroid Disorders	681
AI Applications in Pituitary and Adrenal Disorders	682
Diagnosis and Subtyping	682
Prediction of Treatment Outcomes	682

N. Hong (✉) · Y. Rhee

Department of Internal Medicine, Endocrine Research Institute, Yonsei University College of Medicine, Seoul, South Korea

e-mail: nkhong84@yuhs.ac; YUMIE@yuhs.ac

Y. Park

Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, South Korea

e-mail: YURANGPARK@yuhs.ac

S. C. You

Department of Preventive Medicine, Yonsei University College of Medicine, Seoul, South Korea

© Springer Nature Switzerland AG 2022

N. Lidströmer, H. Ashrafiyan (eds.), *Artificial Intelligence in Medicine*,
https://doi.org/10.1007/978-3-030-64573-1_328

673

Implications of AI for the Endocrinologists	682
References	683

Abstract

This chapter focuses on applications of artificial intelligence (AI) in endocrinology. Endocrinology is the field of medicine that relates to the endocrine system that is consisted of endocrine glands, hormones, target organs, and feedback loop to maintain metabolic homeostasis. Endocrinology covers a broad range of health-related issues from common diseases such as diabetes mellitus, osteoporosis, and hypothyroidism to the rare diseases such as Cushing disease and acromegaly. The principle of endocrine system is based on feedback loop mediated by hormones, which makes it ideal to deploy AI system by both mechanistic (deductive) and statistical (inductive) modeling. In this chapter, we will review the current AI applications in major domains of endocrinology including diabetes mellitus, bone and mineral disorders, thyroid disorders, and pituitary and adrenal disorders. Each domain has unique tasks that may be improved by AI application. Supervised learning is the most commonly used algorithm, with increasing trend to apply unsupervised or reinforcement learning. To apply AI in screening, diagnosis, risk prediction, and treatment decision of endocrine disorders, various data sources (electronic medical records, medical images, laboratory tests) are currently being used in combination or separately. Finally, we conclude with a perspective for endocrinologists regarding current AI applications in endocrinology fields.

Keywords

Endocrine system · Machine learning · Deep learning · Hormone · Diabetes mellitus · Osteoporosis · Digital therapeutics · Artificial pancreas

Introduction

This chapter focuses on applications of artificial intelligence (AI) in endocrinology. Endocrinology encompasses a broad range of studies of endocrine glands, hormones, target organs, and feedback system [1, 2]. The term hormone, derived from a Greek meaning “to set in motion,” describes the dynamic nature of hormones and their actions on regulating physiologic responses through feedback system to maintain metabolic homeostasis. AI can be defined as “the science and engineering of making intelligent machines, especially intelligent computer programs,” based on the definition proposed by John McCarthy [3, 4]. Current AI mainly focuses on learning and reasoning, although more broad concepts of intelligence including self-awareness, introspection, heuristics, and practical knowledge can be encompassed in the concept of AI. Machine learning (ML) can be defined as a subset of AI, that is, an AI technique to design and train software algorithms to learn from and act on data [5]. The endocrine system can be assessed primarily by evaluation of hormone concentration, providing ideal milieu for deploying AI (or ML) that can be integrated instantly into clinical scenario. The inherent nature of hormone as circulating biologic mediator suggests the potential of AI to improve the prediction of complex systemic responses by dysregulation of hormone system. Major domains of endocrinology include diabetes mellitus, bone and mineral disorders, thyroid disorders, and pituitary and adrenal disorders. The number of AI-related publications is increasing exponentially from early 2000 in all domains (Fig. 1).

Each domain requires the active role of AI to solve unmet needs in various aspects of clinical practice including screening, diagnosis, subtyping, risk prediction, and therapeutic response prediction. In this chapter, we will review exemplary

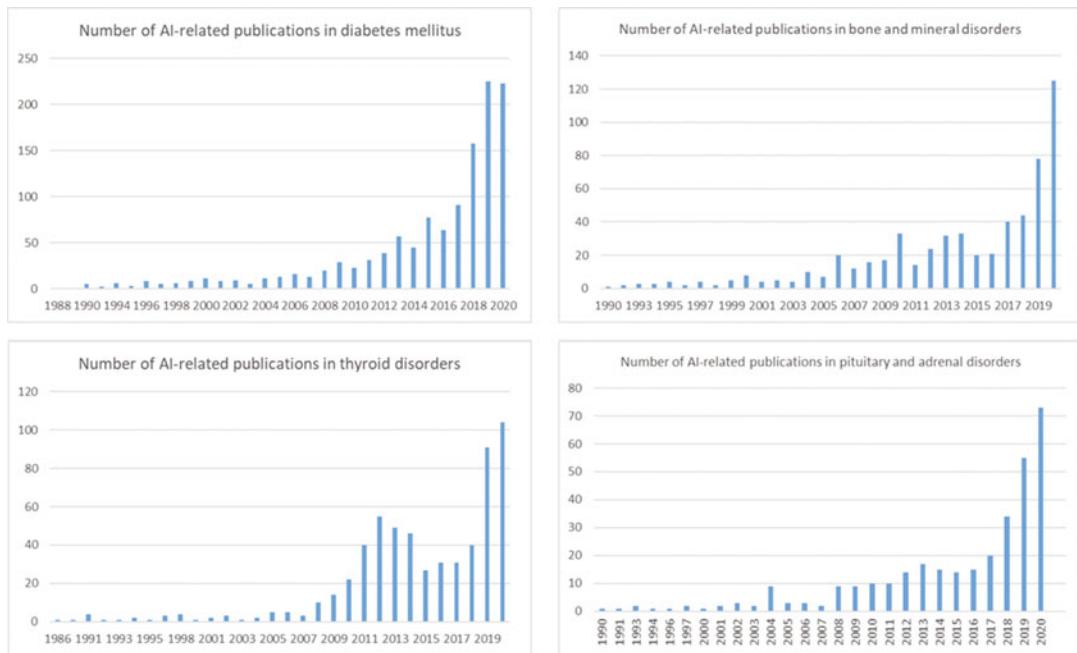


Fig. 1 Number of AI-related publications in major domains of endocrinology (<https://pubmed.ncbi.nlm.nih.gov/>, accessed on Feb 18, 2021)

studies that aimed to deploy AI to tackle the clinical unmet needs in the endocrinology field. Well-performed AI-related studies in endocrinology fields had clearly defined clinical problem and hypothesis in common, with the application of AI algorithms suitable to the characteristics of the problem. Most of studies are based on supervised learning that requires a labeled dataset with mapped output to input dataset to train the model. Recent advances in AI enabled inductive modeling approach to find novel risk factors or clusters by unsupervised learning algorithms, which explores patterns in unlabeled datasets. A few but important studies addressed potential to apply reinforcement learning in optimizing dose adjustment of hormone replacement, which is a learning algorithm to find optimal strategy in unstructured and complex environment to maximize rewards and minimize penalty. Choosing the right metrics to measure model performance is another key step to adequately prove the clinical utility of AI models. Most of studies presented the area under the receiver operating characteristic curve (AUROC, plotting true

positive rate against false positive rate) as the primary performance metric, which describes model classification performance (close to 1 indicates better classifier). Other metrics were also presented together along with AUROC including precision (positive predictive value), recall (sensitivity), accuracy, F1 score, and area under the precision-recall curve (AUPRC) appropriately. Of note, AUPRC may have advantage over AUROC in estimating model performances in an imbalanced dataset, which is a frequently observed case in real-world clinical data [6].

AI Applications in Diabetes Mellitus

AI-Driven Diabetes Care: Changing the Landscape

The role of AI technology is rapidly emerging in diabetes care, with the hope for data-driven precision care [7, 8]. AI-driven changes occur in various domains of diabetes care including the

detection of diabetic complications such as retinopathy, diabetes risk prediction, patient empowering, and lifestyle intervention. The Food and Drug Administration (FDA) provided intensive guidance to company for the development, clinical trial, and review of health-related AI software in 2018, which indicate the rapidly growing healthcare need for AI-driven approach [9]. As reflecting the current trend toward AI-empowered clinical practice in diabetes care, association of Italian diabetologists presented official position statement that AI and new digital technologies can be the key for scientific innovation and accomplishment of precise medicine, if well used [10].

Prediction of Diabetes Risk

Thirty to 80% of patients with prediabetes or diabetes remain undetected worldwide, posing serious gap between disease detection and accurate preventive action [11]. Leveraging accurate prediction of individuals at high risk for diabetes by AI solutions may enable targeted preventive intervention. In a cohort of 852,454 individuals with prediabetes, Cahn and colleagues built machine learning models based on 4.9 million time points using 900 features [12]. The models outperformed logistic regression model in all validation sets, with AUROC above 0.85. Zhang and colleagues reported robust performance of various machine learning models in rural Chinese cohort setting, with plateaued model performance of AUROC 0.87 with 30 clinical variables [13]. In National Health and Nutritional Examination Survey 2013–2014 dataset, combined use of feature selection and machine learning approach increased the predictive performance of prediabetes screening model compared to CDC prediabetes screening tool [14]. In a meta-analysis of 23 studies with 40 prediction models to predict the risk of diabetes, performance of AI models to predict type 2 diabetes was good (pooled c-index 0.812) in the community setting [15]. These studies indicate the potential of AI-driven models as an effective a priori screening tool for individuals at high risk of diabetes onset that can be integrated

into large clinical system [16, 17]. Wearable sensors, Internet-of-thing (IOT)-based monitoring applications, and synthetic data approach showed potential to improve prediction for the diabetes risk in addition to conventional clinical features, although further improvement of sensor accuracy and standardization of collected dataset need to be obtained to realize this approach in prediction of individualized risk of diabetes [18, 19]. Avram and colleagues used smartphone-based photoplethysmography data to create a deep neural network (DNN) model to detect prevalent diabetes [20]. DNN-based score achieved AUROC of 0.830 alongside age, gender, ethnicity, and body mass index, which retained independent predictive value after adjustment for other covariates. This result suggests the feasibility of digital biomarker from smartphone signals, which can be extended to various domains.

Gestational diabetes occurs in 3 to 9% of pregnancies, usually diagnosed at 24–28 weeks of gestation [21]. Because gestational diabetes is associated with adverse outcomes, early intervention to reduce the risk of gestational diabetes is important. To establish an AI model to improve the prediction of gestational diabetes at earlier phase, Artzi and colleagues performed retrospective cohort study using nationwide electronic health record of 588,622 pregnancies from 368,351 women between 2010 and 2017 in Israel [22]. The dataset included demographics, anthropometrics, diagnosis codes, pharmaceutical claims, and laboratory tests. Developed AI model outperformed the standard screening tool (AUROC 0.80 vs. 0.68). More importantly, they derived a simple nine-question self-reportable survey form which may enhance prediction of gestational diabetes even at the initiation of gestation. Improvement of AUPRC was more evident than that of AUROC, providing implication about the importance of comparing AUPRC in imbalanced dataset in real-world setting. Similar approach was successful in developing seven-variable simple model in predicting individuals at high risk for gestational diabetes in Chinese women, with modest to good performance (AUROC 0.77) [23]. While most of AIs trying to predict the risk of gestational diabetes risk are

based on electronic medical record dataset, some researchers tested the performance of AI-based mobile application that showed modest discrimination above AUROC 0.7 using various algorithms, which showed potential to improve current clinical practice with less operation cost and higher efficacy [24, 25].

Retinopathy Detection

Diabetic retinopathy is the most common serious complication of diabetes mellitus leading to vision loss [26]. Early detection of diabetic retinopathy would prevent vision loss by effective intervention modality. However, traditional care models of diabetic retinopathy have failed to ensure access to care, particularly in low-income countries, at least partly due to lack of medical profession to interpret retinal images properly. This clinical major gap calls a novel, innovative approach to screen diabetic retinopathy. Gulshan and colleagues developed retinal image classification model using deep convolutional neural network using the 128,175 retinal images as train dataset. The model showed high sensitivity (>90%), specificity (>90%), and AUROC (>0.9) in two separate, external validation sets [27]. Application of AI for diabetic retinopathy screening is now moving forward to real-world clinical application beyond research level [28]. A group of studies validated deep learning algorithms externally in different countries [29–33]. AI solutions for diabetic retinopathy screening have also cleared regulatory approval in the United States and Europe based on reported performance of AI solutions in clinical trials. The first medical device (IDx-DR), an AI software program that analyzes retinal camera images, enabled on-site screening of diabetic retinopathy by non-eye professionals in primary care settings, which resulted in improved access to standard care and higher patient satisfaction [34]. Furthermore, a group of studies addressed the possibilities of utilizing AI to interpret retinal images from handheld portable fundus camera or smartphone-based retinal imaging, which may have potential to improve the accessibility even further

[35, 36]. In Singapore, the estimated annual savings by adopting semi-automated diabetic retinopathy screening model is projected to be \$15 million by 2050, providing strong rationale for using deep learning system as assistive screening tool for diabetic retinopathy in economic aspects [37].

Prediction of Diabetic Complications

Diabetic complications include microvascular complications (retinopathy, nephropathy, and neuropathy) and macrovascular complications (peripheral artery diseases, coronary artery diseases), which lead to serious morbidity and mortality [38]. Hypoglycemia during diabetes treatment may contribute to increased risk of cardiovascular events. Diabetes is also related to increased risk of pancreatic cancer [39, 40]. Therefore, accurate prediction of the complication risk in patients with diabetes mellitus would enable individualized intervention, which may reduce or prolong the onset of complication, related morbidities, and mortality. In 2287 patients with type 2 diabetes who underwent metabolic surgery matched 1:5 to 11,435 nonsurgical patients, Aminian and colleagues built the machine learning models to predict 10-year risk for all-cause mortality, coronary artery events, heart failure, and nephropathy [41]. All models showed modest to good performance (0.73 to 0.81), with the ability to calculate personalized risk for patients with type 2 diabetes with or without metabolic surgery. Similar approach using electronic medical record type of data also yielded prediction accuracy between 73% and 83% when recurrent neural network models were used, which outperformed the 66% to 76% accuracy of traditional models [42]. Longitudinal electronic health dataset can provide rich features to predict trajectory to diabetic complications when the proper dataset meets the appropriate machine learning tools. In a study, electronic health records were analyzed with temporal-enhanced gradient boosting machine to predict 1-year chronic kidney disease risk (AUROC 0.78 to 0.83), which outperformed other models [43]. Various machine learning

algorithms were applied to create AI models to predict the risk of coronary heart disease, heart failure, diabetic foot, and hypoglycemia [42, 44–49]. Segar and colleagues developed WATCH-DM score to predict the risk of heart failure using data from 8756 patients with diabetes who were enrolled in the Action to Control Cardiovascular Risk in Diabetes (ACCORD) trial [50]. A model based on simple seven variables well predicted the risk of heart failure (C-index 0.74 in external validation set). Most importantly, authors derived easy-to-use integer-based calculator (WATCH-DM score) from machine learning model which can be calculated by hands, with comparable discriminatory performance to that of random survival forest model (C-index 0.70). This study gives unique insights in several aspects. First, post hoc analysis of well-conducted, validated dataset from large randomized clinical trials ensured the high quality of data and annotation of clinical events. Second, creating easy-to-use hand calculator for clinicians provided opportunity to utilize the results from the study even in primary clinic setting without the need to access the machine learning models, although the validity of this score needs to be further tested in various populations.

Continuous Glucose Monitoring and Closed-Loop Artificial Pancreas System

Recent advances in continuous glucose monitoring system (CGMS) and artificial pancreas broadened the therapeutic options for patients with diabetes, which was empowered by advances in cloud computing power [9]. In order to attain desired glucose level while avoiding hypoglycemia in patients with type 1 diabetes, right amount of insulin needs to be delivered subcutaneously at the right time according to circulating blood glucose level, which is one of the main roles of the pancreas as an endocrine organ in human body. In the ADVICE4U trial, a 6-month, multicenter, randomized, non-inferiority trial in 108 participants with type 1 diabetes on insulin pump therapy, the percentage of time spent within the target glucose

range in the AI-based decision support system (AI-DSS) arm was statistically non-inferior to the expert physician arm [51]. Of note, three severe adverse events (two hypoglycemia and one diabetic ketoacidosis) were reported in the physician arm, whereas none were reported in AI-DSS arm. This study reveals the AI-DSS can be non-inferior to expert physician, which may be a breakthrough for individuals with type 1 diabetes on insulin pump who have limited access to specialized clinic. For the automated closed-loop system, the FDA approved artificial pancreas system (Control-IQ) that delivers insulin as needed according to CGMS signals in 2020 [52]. During the long waiting for the results of clinical trials and the FDA approval for closed-loop artificial pancreas system, an interesting phenomenon was observed that patients with diabetes themselves developed the DIYPS (Do It Yourself Pancreas System) and OpenAPS (Open Artificial Pancreas System) using available CGMS, insulin pump, and cloud computing instead of awaiting for the regulatory body approval [53]. Although the DIYPS or OpenAPS is not medically approved or regulated yet, this phenomenon clearly provides an example of rapidly evolving healthcare environment and patient-led healthcare model empowered by AI. At the same time, it raises many ethical and legal considerations and accountability issues that need to be solved between patients, healthcare providers, manufacturers, and regulatory agents. Nevertheless, there is still room for improvement by AI in making better artificial pancreas system, such as more delicate control of bihormonal artificial pancreas delivering both glucagon and insulin according to individual need of patients [54].

Therapeutic Lifestyle Modification

Zeevi and colleagues continuously monitored glucose level after 46,898 meals in a cohort of 800 individuals [55]. They developed an AI algorithm to predict individual response to meals via analyzing multimodal dataset including blood parameters, dietary habits, anthropometrics, physical activity, and gut microbiota. They further

validated the algorithm in randomized clinical trial of 100-person cohort, which revealed significantly lower postprandial responses and alteration in gut microbiome in intervention arm. This study provides ideal example of translating AI-driven knowledge to real-world practice, suggesting that personalized diet recommendation may be key to successfully alter metabolic consequences rather than conventional, fixed recommendation. Digital therapeutics is defined as a new class of digital health that provides health-related intervention with digital technologies (usually apps or Internet-based) [56]. Digital therapeutics can improve self-confidence of patients, which effectively changes the health behavior of patients by tracking daily log, weight and activity monitoring, and providing nutrition guides and feedback leading to decreased hemoglobin A1c and depression [57, 58]. A study estimated that digital therapeutics would save health resource utilization up to \$145 per patient per month, with higher potential benefits in patients with diabetes [59]. Rapidly increasing amount of patient-generated health data may enhance the feasibility and efficacy of digital therapeutics market, although the firm, successful models in digital therapeutics in both clinical efficacy and marketing are still awaited.

AI Applications in Bone and Mineral Disorders

Fracture Identification

Imaging data plays a pivotal role to diagnose fracture. Deep neural network models to detect fracture may assist clinical practice of radiologists with improved accuracy and lessened time-intensive workload in interpreting plain x-ray, computed tomography (CT), or magnetic resonance imaging (MRI) scans [60]. Several studies demonstrated good diagnostic performance of AI models, sometimes achieving discriminatory performance similar to human expert-level performance, using deep learning to detect fracture at various sites [61–64]. Of note, deep learning approaches to computer-aided detection of

fracture significantly improved diagnostic accuracy of wrist fracture from x-ray scans from 81% (unaided) to 91% (aided), with 47% reduction of misinterpretation rate in average clinician [61]. The FDA has cleared a deep learning, AI software to help in the identification of wrist fracture in adults (OsteoDetect, Imagen) in May 2018, showing innovative changes in premarket review pathway of regulatory bodies to accelerate the development of AI to aid in clinical practice [65]. Fracture liaison service, a program to prevent re-fracture after initial fracture, can be a feasible target to integrate image-based fracture detection in clinical practice [66]. Radiology reports can be a good source to enhance fracture detection by AI utilization in the setting of fracture liaison service. White and colleagues developed XRAIT, a natural language processing algorithm to detect fracture from radiology reports, which showed better performance in detection of fracture cases than manual finding [67].

Opportunistic Screening of Osteoporosis and Sarcopenia

Proper detection of osteoporosis, a status of low bone strength with impaired bone quantity and quality, is important to detect individuals at high risk of fragility fracture. Dual-energy x-ray absorptiometry (DXA) is the current standard tool to diagnose osteoporosis, defined as low bone mineral density (BMD, T score – 2.5 or lower) at the spine or hip [68, 69]. However, limited accessibility to DXA machine in some regions is one reason that contribute to current underdiagnosis of osteoporosis [70]. AI-driven opportunistic screening of osteoporosis from various data sources (mainly plain x-ray or CT) can be a solution to optimize the diagnosis of osteoporosis. In studies of Pickhardt and a group of researchers, Hounsfield unit (HU) of trabecular bone region in L1 vertebra obtained from routine clinical CT scans showed linear decreasing trajectory across age, with the potential thresholds within the range of 90 to 135 HU corresponding to DXA-defined osteoporosis [71]. Importantly, the measurement of L1 HU was performed fully

automatically with high agreement with manual assessment, suggesting the feasibility of large-scale application for the screening of osteoporosis in clinical practice. Features obtained from radiomics approach also have potential to improve osteoporosis detection in routine clinical CT scans [72]. Vertebral and femoral strength can also be analyzed with finite element modeling from routine abdominal CT, which showed good agreement with DXA-derived areal BMD [73]. AI application in CT images holds promise for detecting impaired muscle mass and function (sarcopenia), which is another potential therapeutic target to prevent falls leading to incident fracture [74, 75]. Plain x-ray can be a source to predict DXA-equivalent BMD, although several issues in preprocessing of image quality and heterogeneity in scan protocol need to be overcome [76]. AI-driven opportunistic screening of osteoporosis from dental panoramic images has been developed to improve the diagnosis rate of osteoporosis, which merit further investigation to improve accuracy with external validation in large scale [77].

Fracture Risk Assessment

Accurate fracture risk assessment is the key step to initiate personalized therapeutic interventions to prevent fracture including pharmacologic and non-pharmacologic strategies [69, 78]. There are well-validated fracture risk estimation models, such as FRAX, based on clinical predictors with or without BMD values, which served well to identify individuals at high risk of fracture [78]. Alongside the current standard clinical models, recent studies are seeking novel AI-based algorithms to enhance fracture risk prediction using various data sources including electronic health records, imaging, or omics dataset. Leslie and colleagues developed convolutional neural network algorithm to detect vertebral fracture in lateral spine images obtained from DXA [79]. Developed deep learning algorithm well detected the vertebral fracture with AUROC (0.94) comparable to that of expert, with additive prognostic value for incident nonvertebral

fracture independent of baseline FRAX probability. Although several AI models showed potential to improve fracture prediction based on structured clinical variables, artificial intelligence may have wider application in unstructured data sources rather than structured dataset [80–84]. In a small proof-of-concept study using DXA images, patient-specific finite element modeling with machine learning application showed potential to outperform standard DXA BMD in prediction of fracture [85]. In a community-based older men cohort (MrOS cohort), genomic risk score calculated from 1103 associated single nucleotide polymorphisms showed modest to good discrimination for incident fracture along with clinical variables, although the feature importance of genomic risk score was lower compared to conventional clinical predictors such as BMD, age, and weight [86]. Natural language processing can be useful to develop fracture prediction model based on electronic health record data. Cummings and colleagues developed an AI based on natural language processing, Crystal Bone, to predict short-term fracture risk within 1 to 2 years from electronic health record data of over one million patients [87]. The model predicted short-term fracture risk with high accuracy (AUROC 0.81), with significant improvement of identifying high-risk individuals when compared to human performance retrospectively. These findings suggest the potential role of AI in opening the new opportunity to improve fracture prediction by utilizing unstructured features along with well-established structured clinical predictors, which need to be validated in future studies.

Finding Novel Biomarkers Related to Bone Metabolism

AI-driven approach can have unique strength in identifying novel features correlated to bone metabolism in complex, unstructured dataset, potentially leading to identification of new biomarkers or biologic pathway. AI application enabled the integration of transcriptome-wide association study and genome-wide association study that revealed novel candidate genes associated with

osteoporosis [88]. High-throughput sequencing of transcriptomic or proteomic profiles in human, powered by bioinformatics, may reveal differentially expressed RNA profiles and protein networks related to alteration in bone and muscle metabolism, leading to the discovery of novel diagnostic or therapeutic target [89, 90].

AI Applications in Thyroid Disorders

AI Application in Thyroid Cancer

Diagnosis of thyroid cancer is frequently performed by thyroid ultrasonography. In a large scale, retrospective multicohort diagnostic study performed in China, deep convolutional neural network model improved specificity in identifying patients with thyroid cancer (88% vs. 69%) compared to radiology experts while maintaining similar sensitivity (85% vs. 89%) in an external validation set [91]. Of note, machine learning-assisted ultrasonographic visual approach showed better diagnostic performance than ultrasonography alone (AUROC 0.95 vs. 0.92), with significant reduction of unnecessary invasive needle aspiration procedure from 30% to 5% compared to current standard risk stratification system [92]. Radiologic features of thyroid nodules in ultrasonography may also convey information of genetic risk. A study found that automated machine learning applied to ultrasonography were able to detect thyroid nodules with high-risk mutations on molecular testing, which showed possibility for the diagnostic application of AI for predicting noninvasive high-risk mutation in thyroid nodule [93]. The diagnosis can remain indeterminate even after cytologic examination by fine needle aspiration of thyroid nodule. Machine learning algorithms based on next-generation RNA sequencing data obtained from cytologic examination were able to discriminate benign nodule in cytologically indeterminate thyroid nodules [94]. The algorithm showed high sensitivity and accuracy for identifying benign nodules with increased specificity, which potentially reduces unnecessary diagnostic surgery. As the frequency of fine needle aspiration procedure

is increasing along with increasing incidence of thyroid nodule, the role of AI is also expanding in interpreting thyroid pathology, particularly for indeterminate pathologic findings. A systematic review of 19 studies revealed modest to high correlation between AI-aided diagnosis and expert pathologist diagnosis, suggesting promising results for AI application in thyroid pathology if technical issues such as computational burden due to a large size of whole slide images can be resolved [95]. Although determining lymph node metastasis risk in papillary thyroid cancer is the key information that guides diagnosis and treatment process, the performance of lymph node metastasis detection by ultrasonography remains suboptimal. In a retrospective cohort of 3172 patients, transfer learning radiomics model enhanced the prediction for lymph node metastasis at preoperative phase, which outperformed clinical statistical model [96]. These findings provide good examples on how AI technology can drive clinical practices in managing thyroid cancer toward enhanced patient safety and higher diagnostic accuracy.

AI Application in Functional Thyroid Disorders

Graves' disease is a common autoimmune disorder resulting in hyperthyroidism. Although anti-thyroidal drugs, radioactive iodine, or surgery can effectively control the excessive production of thyroid hormones in some patients, the therapeutic options for severe, refractory, or recurrent cases are still limited. A study of key gene co-expression modules and functional pathways in Graves' diseases revealed the pivotal role of bioinformatics and AI in providing novel insights of pathogenesis of Graves' disease [97]. Neural network modeling using clinical features of patients with Graves' disease were able to predict recurrence of disease within 2 years of anti-thyroid drug withdrawal [98]. When total or completion thyroidectomy is performed due to various etiologies, synthetic thyroid hormone replacement is inevitable, but the scheme for dose adjustment remains suboptimal. In a study that analyzed

598 patients who underwent total thyroidectomy due to benign thyroid disorders, machine learning model with Poisson regression outperformed conventional weight-based dosing (correctly predicting dose in 64.8% vs. 27.4%) [99]. These studies illustrate the potential of AI application to improve quality of care in functional thyroid disorders.

AI Applications in Pituitary and Adrenal Disorders

Diagnosis and Subtyping

Some endocrinologic disorders, such as hypercortisolism or acromegaly (excessive production of growth hormone), have unique facial characteristics which can have diagnostic value. Usually the recognition of facial characteristics was the role of expert endocrinologists; however, AI principle has the potential to aid in “facial diagnosis” in clinical scenario. In a group of studies, deep neural network algorithms have shown successful results in discriminating patients with acromegaly or hypercortisolism (Cushing’s syndrome) in various dataset [100–102]. MR images play a key role in the diagnosis and subtyping of pituitary tumors, which guide following evaluation and treatment plans. In related MRI-based AI studies, majority of machine learning tasks were classification of nonfunctioning pituitary adenoma versus functioning tumor and prediction of tumor proliferation index [103, 104]. Machine learning-based metabolomics analysis has potential to improve diagnosis and subtyping of endocrine disorders, providing opportunities to explore novel biologic mechanisms and biomarkers [105]. Plasma or urine steroid metabolites revealed a fingerprint of adrenal cortical carcinoma, which predicted the recurrence of cancer with high accuracy [106, 107]. By combining metabolomics with immunohistochemistry data of pheochromocytoma tumor tissue, genetic alteration with prognostic value, such as succinate dehydrogenase mutational status, could be predicted [108]. Well-established clinical features can be a good data source to create a useful machine learning model. Mulatero and colleagues

designed a machine learning-based score to predict subtype of primary aldosteronism using clinical features that are collected during standard clinical practice [109]. This model may have potential to guide surgical decision in centers where the gold standard test adrenal vein sampling, an invasive, technically demanding procedure, is unavailable.

Prediction of Treatment Outcomes

In patients with pituitary or adrenal tumor, it is important to predict clinical outcome of surgical or medical treatment to support therapeutic decision. This is particularly the case when the surgical procedure accompanies high risk of complications, such as hypophysectomy in pituitary tumor. Machine learning approach based on image or clinical data has shown several remarkable results in predicting treatment outcomes in patients with pituitary tumors such as acromegaly or Cushing’s disease [110–113]. MRI alone provided high discriminatory performance (AUROC 0.85) for classifying responders to drug treatment in patients with acromegaly [114]. Similarly, machine learning models using CT-derived texture features of adrenal metastatic masses predicted tumor progression and survival, revealing potential of image data as a window to biologic heterogeneity in endocrine tumors [115].

Implications of AI for the Endocrinologists

In this chapter, we reviewed current applications of AI in the field of endocrinology. As Eric J. Topol clearly addressed in his article, the narrative of bringing AI to medicine just has begun, with hope for improving workflow, reducing medical errors, and empowering patients [116]. Given the current trend of exponentially increasing publications regarding AI applications in various domains of endocrinology, this may also hold true for endocrinology sector. Application of AI in endocrinology has potential to transform current clinical practice to more connected care (Fig. 2). However, clinical gestalt of endocrinologists should play a

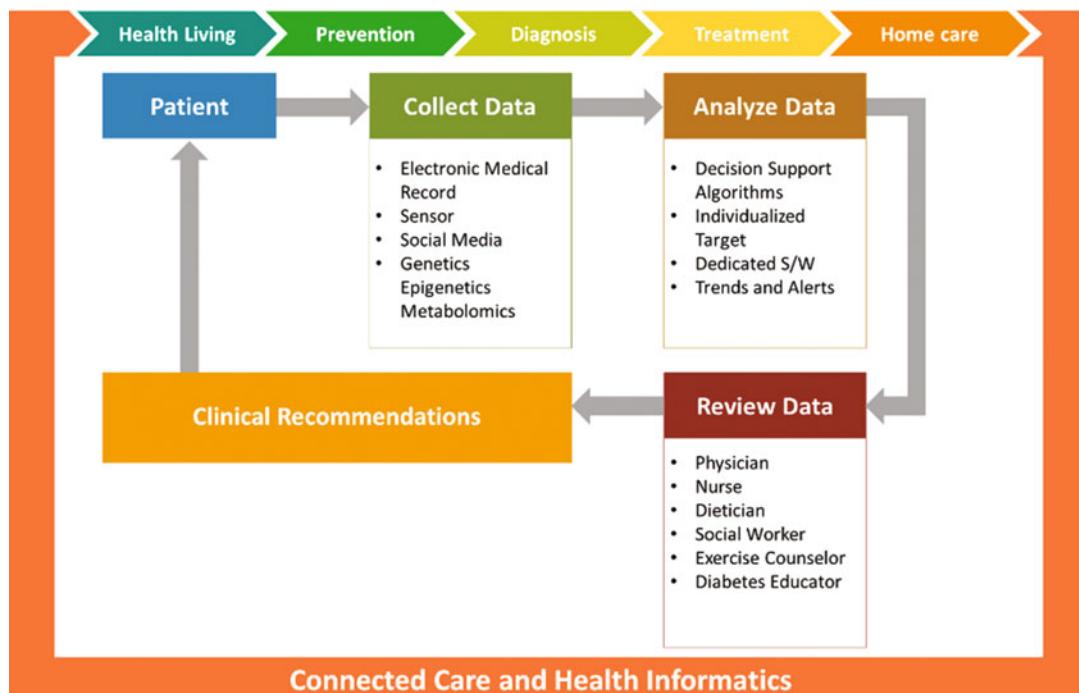


Fig. 2 Transforming current endocrinology practice to more connected healthcare using AI application

pivotal role in evaluating whether the results from AI models are believable, with scrutiny and scientific rigor. While AI is finding its way to the integration into clinical application in endocrinology, endocrinologists need to learn how to understand and use AI, with healthy skepticism [117]. In clinical perspectives, moving toward the application of precision medicine in endocrinology requires high-performance AI models that are well validated, easy to access, and easy to use [118].

References

- Jameson JL. Approach to the patient with endocrine disorders. In: Jameson JL, Fauci AS, Kasper DL, Hauser SL, Longo DL, Loscalzo J, editors. Harrison's principles of internal medicine. 20th ed. New York: McGraw-Hill Education; 2018.
- Molina PE. Chapter 1. General principles of endocrine physiology. In: Endocrine physiology. 4th ed. - New York: McGraw-Hill; 2013.
- McCarthy J. What is artificial intelligence? Personal website. 2007. <http://www-formal.stanford.edu/jmc/>
- McCarthy J. From here to human-level AI. Artif Intell. 2007;171(18):1174–82. <https://doi.org/10.1016/j.artint.2007.10.009>.
- Beaulieu-Jones B, Finlayson SG, Chivers C, Chen I, McDermott M, Kandola J, Dalca AV, Beam A, Fiterau M, Naumann T. Trends and focus of machine learning applications for health research. JAMA Netw Open. 2019;2(10):e1914051. <https://doi.org/10.1001/jamanetworkopen.2019.14051>.
- Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432. <https://doi.org/10.1371/journal.pone.0118432>.
- Ellahham S. Artificial intelligence: the future for diabetes care. Am J Med. 2020;133(8):895–900. <https://doi.org/10.1016/j.amjmed.2020.03.033>.
- Vu GT, Tran BX, McIntyre RS, Pham HQ, Phan HT, Ha GH, Gwee KK, Latkin CA, Ho RCM, Ho CSH. Modeling the research landscapes of artificial intelligence applications in diabetes (GAP(RESEARCH)). Int J Environ Res Public Health. 2020;17(6):1982. <https://doi.org/10.3390/ijerph17061982>.
- Broome DT, Hilton CB, Mehta N. Policy implications of artificial intelligence and machine learning in diabetes management. Curr Diab Rep. 2020;20(2):5. <https://doi.org/10.1007/s11892-020-1287-2>.
- Musacchio N, Giancaterini A, Guaita G, Ozzello A, Pellegrini MA, Ponzani P, Russo GT, Zilich R, de Michelis A. Artificial intelligence and big data in diabetes care: a position statement of the Italian Association of Medical Diabetologists. J Med Internet Res. 2020;22(6):e16922. <https://doi.org/10.2196/16922>.

11. Brown N, Critchley J, Bogowicz P, Mayige M, Unwin N. Risk scores based on self-reported or available clinical data to detect undiagnosed type 2 diabetes: a systematic review. *Diabetes Res Clin Pract.* 2012;98(3):369–85. <https://doi.org/10.1016/j.diabres.2012.09.005>.
12. Cahn A, Shoshan A, Sagiv T, Yesharim R, Goshen R, Shalev V, Raz I. Prediction of progression from pre-diabetes to diabetes: development and validation of a machine learning model. *Diabetes Metab Res Rev.* 2020;36(2):e3252. <https://doi.org/10.1002/dmrr.3252>.
13. Zhang L, Wang Y, Niu M, Wang C, Wang Z. Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: the Henan Rural Cohort Study. *Sci Rep.* 2020;10(1):4406. <https://doi.org/10.1038/s41598-020-61123-x>.
14. De Silva K, Jonsson D, Demmer RT. A combined strategy of feature selection and machine learning to identify predictors of prediabetes. *J Am Med Inform Assoc.* 2019;27:396. <https://doi.org/10.1093/jamia/ocz204>.
15. Silva K, Lee WK, Forbes A, Demmer RT, Barton C, Enticott J. Use and performance of machine learning models for type 2 diabetes prediction in community settings: a systematic review and meta-analysis. *Int J Med Inform.* 2020;143:104268. <https://doi.org/10.1016/j.ijmedinf.2020.104268>.
16. Bernardini M, Romeo L, Misericordia P, Frontoni E. Discovering the type 2 diabetes in electronic health records using the sparse balanced support vector machine. *IEEE J Biomed Health Inform.* 2020;24(1):235–46. <https://doi.org/10.1109/jbhi.2019.2899218>.
17. Kopitar L, Kocbek P, Cilar L, Sheikh A, Stiglic G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Sci Rep.* 2020;10(1):11981. <https://doi.org/10.1038/s41598-020-68771-z>.
18. Baig MM, GholamHosseini H, Gutierrez J, Ullah E, Lindén M. Early detection of prediabetes and T2DM using wearable sensors and internet-of-things-based monitoring applications. *Appl Clin Inform.* 2021;12(1):1–9. <https://doi.org/10.1055/s-0040-1719043>.
19. Stolfi P, Valentini I, Palumbo MC, Tieri P, Grignolio A, Castiglione F. Potential predictors of type-2 diabetes risk: machine learning, synthetic data and wearable health devices. *BMC Bioinformatics.* 2020;21(Suppl 17):508. <https://doi.org/10.1186/s12859-020-03763-4>.
20. Avram R, Olgin JE, Kuhar P, Hughes JW, Marcus GM, Pletcher MJ, Aschbacher K, Tison GH. A digital biomarker of diabetes from smartphone-based vascular signals. *Nat Med.* 2020;26(10):1576–82. <https://doi.org/10.1038/s41591-020-1010-5>.
21. American Diabetes Association. 2. Classification and diagnosis of diabetes: standards of medical care in diabetes – 2018. *Diabetes Care.* 2018;41(Suppl 1): S13–27. <https://doi.org/10.2337/dc18-S002>.
22. Artzi NS, Shilo S, Hadar E, Rossman H, Barlash-Hazan S, Ben-Haroush A, Balicer RD, Feldman B, Wiznitzer A, Segal E. Prediction of gestational diabetes based on nationwide electronic health records. *Nat Med.* 2020;26(1):71–6. <https://doi.org/10.1038/s41591-019-0724-8>.
23. Wu YT, Zhang CJ, Mol BW, et al. Early prediction of gestational diabetes mellitus in the Chinese population via advanced machine learning. *J Clin Endocrinol Metab.* 2020;106:e1191. <https://doi.org/10.1210/clinem/dgaa899>.
24. Shen J, Chen J, Zheng Z, et al. An innovative artificial intelligence-based app for the diagnosis of gestational diabetes mellitus (GDM-AI): development study. *J Med Internet Res.* 2020;22(9):e21573. <https://doi.org/10.2196/21573>.
25. Albert L, Capel I, Garcia-Saez G, Martin-Redondo P, Hernando ME, Rigla M. Managing gestational diabetes mellitus using a smartphone application with artificial intelligence (SineDie) during the COVID-19 pandemic: much more than just telemedicine. *Diabetes Res Clin Pract.* 2020;169:108396. <https://doi.org/10.1016/j.diabres.2020.108396>.
26. Leasher JL, Bourne RR, Flaxman SR, et al. Global estimates on the number of people blind or visually impaired by diabetic retinopathy: a meta-analysis from 1990 to 2010. *Diabetes Care.* 2016;39(9): 1643–9. <https://doi.org/10.2337/dc15-2171>.
27. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA.* 2016;316(22):2402–10. <https://doi.org/10.1001/jama.2016.17216>.
28. Gunasekeran DV, Ting DSW, Tan GSW, Wong TY. Artificial intelligence for diabetic retinopathy screening, prediction and management. *Curr Opin Ophthalmol.* 2020;31(5):357.
29. Bellemo V, Lim G, Rim TH, et al. Artificial intelligence screening for diabetic retinopathy: the real-world emerging application. *Curr Diab Rep.* 2019;19(9):72. <https://doi.org/10.1007/s11892-019-1189-3>.
30. Bellemo V, Lim ZW, Lim G, et al. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. *Lancet Digit Health.* 2019;1(1):e35–44. [https://doi.org/10.1016/s2589-7500\(19\)30004-4](https://doi.org/10.1016/s2589-7500(19)30004-4).
31. Gulshan V, Rajan RP, Widner K, et al. Performance of a deep-learning algorithm vs manual grading for detecting diabetic retinopathy in India. *JAMA Ophthalmol.* 2019;137(9):987–93. <https://doi.org/10.1001/jamaophthalmol.2019.2004>.
32. Ruamviboonsuk P, Krause J, Chotcomwongse P, et al. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide

- screening program. *npj Digit Med.* 2019;2(1):25. <https://doi.org/10.1038/s41746-019-0099-8>.
33. Son J, Shin JY, Kim HD, Jung KH, Park KH, Park SJ. Development and validation of deep learning models for screening multiple abnormal findings in retinal fundus images. *Ophthalmology.* 2020;127(1): 85–94. <https://doi.org/10.1016/j.ophtha.2019.05.029>.
34. Keel S, Lee PY, Scheetz J, Li Z, Kotowicz MA, MacIsaac RJ, He M. Feasibility and patient acceptability of a novel artificial intelligence-based screening model for diabetic retinopathy at endocrinology outpatient services: a pilot study. *Sci Rep.* 2018;8(1): 4330. <https://doi.org/10.1038/s41598-018-22612-2>.
35. Rogers TW, Gonzalez-Bueno J, Garcia Franco R, Lopez Star E, Méndez Marín D, Vassallo J, Lanssing VC, Trikha S, Jaccard N. Evaluation of an AI system for the detection of diabetic retinopathy from images captured with a handheld portable fundus camera: the MAILOR AI study. *Eye.* 2021;35(2):632–8. <https://doi.org/10.1038/s41433-020-0927-8>.
36. Karakaya M, Hacisoftwareoglu RE. Comparison of smartphone-based retinal imaging systems for diabetic retinopathy detection using deep learning. *BMC Bioinformatics.* 2020;21(Suppl 4):259. <https://doi.org/10.1186/s12859-020-03587-2>.
37. Xie Y, Nguyen QD, Hamzah H, et al. Artificial intelligence for teleophthalmology-based diabetic retinopathy screening in a national programme: an economic analysis modelling study. *Lancet Digit Health.* 2020;2(5):e240–9. [https://doi.org/10.1016/s2589-7500\(20\)30060-1](https://doi.org/10.1016/s2589-7500(20)30060-1).
38. Papapetodorou K, Banach M, Bekiari E, Rizzo M, Edmonds M. Complications of diabetes 2017. *J Diabetes Res.* 2018;2018:3086167. <https://doi.org/10.1155/2018/3086167>.
39. De Souza A, Irfan K, Masud F, Saif MW. Diabetes type 2 and pancreatic cancer: a history unfolding. *JOP: J Pancreas.* 2016;17(2):144–8.
40. Pereira SP, Oldfield L, Ney A, et al. Early detection of pancreatic cancer. *Lancet Gastroenterol Hepatol.* 2020;5(7):698–710. [https://doi.org/10.1016/s2468-1253\(19\)30416-9](https://doi.org/10.1016/s2468-1253(19)30416-9).
41. Aminian A, Zajicek A, Arterburn DE, Wolski KE, Brethauer SA, Schauer PR, Nissen SE, Kattan MW. Predicting 10-year risk of end-organ complications of type 2 diabetes with and without metabolic surgery: a machine learning approach. *Diabetes Care.* 2020;43(4):852–9. <https://doi.org/10.2337/dc19-2057>.
42. Ljubic B, Hai AA, Stanojevic M, Diaz W, Polimac D, Pavlovski M, Obradovic Z. Predicting complications of diabetes mellitus using advanced machine learning algorithms. *J Am Med Inform Assoc.* 2020;27(9): 1343–51. <https://doi.org/10.1093/jamia/ocaa120>.
43. Song X, Waitman LR, Yu AS, Robbins DC, Hu Y, Liu M. Longitudinal risk prediction of chronic kidney disease in diabetic patients using a temporal-enhanced gradient boosting machine: retrospective cohort study. *JMIR Med Inform.* 2020;8(1):e15510. <https://doi.org/10.2196/15510>.
44. Kodama S, Fujihara K, Shiozaki H, et al. Ability of current machine learning algorithms to predict and detect hypoglycemia in patients with diabetes mellitus: meta-analysis. *JMIR Diabetes.* 2021;6(1): e22458. <https://doi.org/10.2196/22458>.
45. Elhadd T, Mall R, Bashir M, Palotti J, Fernandez-Luque L, Farooq F, Mohanadi DA, Dabbous Z, Malik RA, Abou-Samra AB. Artificial intelligence (AI) based machine learning models predict glucose variability and hypoglycaemia risk in patients with type 2 diabetes on a multiple drug regimen who fast during ramadan (the PROFAST – IT Ramadan study). *Diabetes Res Clin Pract.* 2020;169:108388. <https://doi.org/10.1016/j.diabres.2020.108388>.
46. Yamada T, Iwasaki K, Maedera S, Ito K, Takeshima T, Noma H, Shojima N. Myocardial infarction in type 2 diabetes using sodium-glucose co-transporter-2 inhibitors, dipeptidyl peptidase-4 inhibitors or glucagon-like peptide-1 receptor agonists: proportional hazards analysis by deep neural network based machine learning. *Curr Med Res Opin.* 2020;36(3):403–9. <https://doi.org/10.1080/03007995.2019.1706043>.
47. Cruz-Vega I, Hernandez-Contreras D, Peregrina-Barreto H, Rangel-Magdaleno JJ, Ramirez-Cortes JM. Deep learning classification for diabetic foot thermograms. *Sensors (Basel).* 2020;20(6):1762. <https://doi.org/10.3390/s20061762>.
48. Ferreira A, Ferreira DD, Oliveira HC, Resende IC, Anjos A, Lopes M. Competitive neural layer-based method to identify people with high risk for diabetic foot. *Comput Biol Med.* 2020;120:103744. <https://doi.org/10.1016/j.combiomed.2020.103744>.
49. Fan R, Zhang N, Yang L, Ke J, Zhao D, Cui Q. AI-based prediction for the risk of coronary heart disease among patients with type 2 diabetes mellitus. *Sci Rep.* 2020;10(1):14457. <https://doi.org/10.1038/s41598-020-71321-2>.
50. Segar MW, Vaduganathan M, Patel KV, et al. Machine learning to predict the risk of incident heart failure hospitalization among patients with diabetes: the WATCH-DM risk score. *Diabetes Care.* 2019;42(12): 2298–306. <https://doi.org/10.2337/dc19-0587>.
51. Nimri R, Battelino T, Laffel LM, Slover RH, Schatz D, Weinzimer SA, Dovc K, Danne T, Phillip M. Insulin dose optimization using an automated artificial intelligence-based decision support system in youths with type 1 diabetes. *Nat Med.* 2020;26(9): 1380–4. <https://doi.org/10.1038/s41591-020-1045-7>.
52. Brown SA, Kovatchev BP, Raghinaru D, et al. Six-month randomized, multicenter trial of closed-loop control in type 1 diabetes. *N Engl J Med.* 2019;381(18):1707–17. <https://doi.org/10.1056/NEJMoa1907863>.
53. Jennings P, Hussain S. Do-it-yourself artificial pancreas systems: a review of the emerging evidence and

- insights for healthcare professionals. *J Diabetes Sci Technol.* 2020;14(5):868–77. <https://doi.org/10.1177/1932296819894296>.
54. El-Khatib FH, Balliro C, Hillard MA, et al. Home use of a bihormonal bionic pancreas versus insulin pump therapy in adults with type 1 diabetes: a multicentre randomised crossover trial. *Lancet.* 2017;389(10067):369–80. [https://doi.org/10.1016/s0140-6736\(16\)32567-3](https://doi.org/10.1016/s0140-6736(16)32567-3).
55. Zeevi D, Korem T, Zmora N, et al. Personalized nutrition by prediction of glycemic responses. *Cell.* 2015;163(5):1079–94. <https://doi.org/10.1016/j.cell.2015.11.001>.
56. Kaufman N. Digital therapeutics: leading the way to improved outcomes for people with diabetes. *Diabetes Spectr.* 2019;32(4):301–3. <https://doi.org/10.2337/ds19-0012>.
57. Salazar P, Somauroo A. Chapter 18 – Are digital therapeutics poised to become mainstream in diabetes care? In: Klonoff DC, Kerr D, Mulvaney SA, editors. *Diabetes digital health.* Elsevier; 2020. p. 243–52.
58. Berman MA, Guthrie NL, Edwards KL, Appelbaum KJ, Njike VY, Eisenberg DM, Katz DL. Change in glycemic control with use of a digital therapeutic in adults with type 2 diabetes: cohort study. *JMIR Diabetes.* 2018;3(1):e4. <https://doi.org/10.2196/diabetes.9591>.
59. Nordyke RJ, Appelbaum K, Berman MA. Estimating the impact of novel digital therapeutics in type 2 diabetes and hypertension: health economic analysis. *J Med Internet Res.* 2019;21(10):e15814. <https://doi.org/10.2196/15814>.
60. Kalmet PHS, Sanduleanu S, Primakov S, Wu G, Jochems A, Refaee T, Ibrahim A, Hulst LV, Lambin P, Poeze M. Deep learning in fracture detection: a narrative review. *Acta Orthop.* 2020;91(2):215–20. <https://doi.org/10.1080/17453674.2019.1711323>.
61. Lindsey R, Daluiski A, Chopra S, et al. Deep neural network improves fracture detection by clinicians. *Proc Natl Acad Sci U S A.* 2018;115(45):11591–6. <https://doi.org/10.1073/pnas.1806905115>.
62. Tomita N, Cheung YY, Hassanpour S. Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. *Comput Biol Med.* 2018;98:8–15. <https://doi.org/10.1016/j.combiomed.2018.05.011>.
63. Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, Endo N. Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. *Skelet Radiol.* 2019;48(2):239–44. <https://doi.org/10.1007/s00256-018-3016-3>.
64. Zhang B, Jia C, Wu R, Lv B, Li B, Li F, Du G, Sun Z, Li X. Improving rib fracture detection accuracy and reading efficiency with deep learning-based detection software: a clinical evaluation. *Br J Radiol.* 2021;94(1118):20200870. <https://doi.org/10.1259/bjr.20200870>.
65. Ratner M. FDA backs clinician-free AI imaging diagnostic tools. *Nat Biotechnol.* 2018;36(8):673–4. <https://doi.org/10.1038/nbt0818-673a>.
66. Ong T, Copeland R, Thiam CN, Cerda Mas G, Marshall L, Sahota O. Integration of a vertebral fracture identification service into a fracture liaison service: a quality improvement project. *Osteoporos Int.* 2020;32:921. <https://doi.org/10.1007/s00198-020-05710-8>.
67. Kolanu N, Brown AS, Beech A, Center JR, White CP. Natural language processing of radiology reports for the identification of patients with fracture. *Arch Osteoporos.* 2021;16(1):6. <https://doi.org/10.1007/s11657-020-00859-5>.
68. Lewiecki EM, Watts NB, McClung MR, Petak SM, Bachrach LK, Shepherd JA, Downs RW Jr. Official positions of the international society for clinical densitometry. *J Clin Endocrinol Metab.* 2004;89(8):3651–5. <https://doi.org/10.1210/jc.2004-0124>.
69. Siris ES, Adler R, Bilezikian J, et al. The clinical diagnosis of osteoporosis: a position statement from the National Bone Health Alliance Working Group. *Osteoporos Int.* 2014;25(5):1439–43. <https://doi.org/10.1007/s00198-014-2655-z>.
70. Lems WF, Raterman HG. Critical issues and current challenges in osteoporosis and fracture prevention. An overview of unmet needs. *Ther Adv Musculoskeletal Dis.* 2017;9(12):299–316. <https://doi.org/10.1177/1759720X17732562>.
71. Jang S, Graffy PM, Ziemilewicz TJ, Lee SJ, Summers RM, Pickhardt PJ. Opportunistic osteoporosis screening at routine abdominal and thoracic CT: normative L1 trabecular attenuation values in more than 20 000 adults. *Radiology.* 2019;291(2):360–7. <https://doi.org/10.1148/radiol.2019181648>.
72. Valentinitisch A, Trebeschi S, Kaesmacher J, Lorenz C, Löffler MT, Zimmer C, Baum T, Kirschke JS. Opportunistic osteoporosis screening in multi-detector CT images via local classification of textures. *Osteoporos Int.* 2019;30(6):1275–85. <https://doi.org/10.1007/s00198-019-04910-1>.
73. Hong N, Lee DC, Khosla S, Keaveny TM, Rhee Y. Comparison of vertebral and femoral strength between White and Asian adults using finite element analysis of computed tomography scans. *J Bone Miner Res.* 2020;35(12):2345–54. <https://doi.org/10.1002/jbm.4149>.
74. Choi H, Hong N, Park N, Kim CO, Kim HC, Choi JY, Youn Y, Rhee Y. Computed tomography-derived skeletal muscle radiodensity predicts peak weight-corrected jump power in older adults: the Korean Urban Rural Elderly (KURE) Study. *Calcif Tissue Int.* 2021;108:764. <https://doi.org/10.1007/s00223-021-00812-9>.
75. Boutin RD, Lenchik L. Value-added opportunistic CT: insights into osteoporosis and sarcopenia. *Am J Roentgenol.* 2020;215(3):582–94. <https://doi.org/10.2214/AJR.20.22874>.

76. Lee S, Choe EK, Kang HY, Yoon JW, Kim HS. The exploration of feature extraction and machine learning for predicting bone density from simple spine X-ray images in a Korean population. *Skelet Radiol.* 2020;49(4):613–8. <https://doi.org/10.1007/s00256-019-03342-6>.
77. Lee JS, Adhikari S, Liu L, Jeong HG, Kim H, Yoon SJ. Osteoporosis detection in panoramic radiographs using a deep convolutional neural network-based computer-assisted diagnosis system: a preliminary study. *Dentomaxillofac Radiol.* 2019;48(1):20170344. <https://doi.org/10.1259/dmfr.20170344>.
78. Kanis JA, McCloskey EV, Johansson H, Oden A, Ström O, Borgström F. Development and use of FRAX in osteoporosis. *Osteoporos Int.* 2010;21(Suppl 2):S407–13. <https://doi.org/10.1007/s00198-010-1253-y>.
79. Derkatch S, Kirby C, Kimelman D, Jozani MJ, Davidson JM, Leslie WD. Identification of vertebral fractures by convolutional neural networks to predict nonvertebral and hip fractures: a registry-based cohort study of dual x-ray absorptiometry. *Radiology.* 2019;293(2):405–11. <https://doi.org/10.1148/radiol.2019190201>.
80. Atkinson EJ, Therneau TM, Melton LJ 3rd, Camp JJ, Achenbach SJ, Amin S, Khosla S. Assessing fracture risk using gradient boosting machine (GBM) models. *J Bone Miner Res.* 2012;27(6):1397–404. <https://doi.org/10.1002/jbmr.1577>.
81. Kruse C, Eiken P, Vestergaard P. Machine learning principles can improve hip fracture prediction. *Calcif Tissue Int.* 2017;100(4):348–60. <https://doi.org/10.1007/s00223-017-0238-7>.
82. Kruse C, Eiken P, Vestergaard P. Clinical fracture risk evaluated by hierarchical agglomerative clustering. *Osteoporos Int.* 2016;28(3):819–32. <https://doi.org/10.1007/s00198-016-3828-8>.
83. de Vries BCS, Hegeman JH, Nijmeijer W, Geerdink J, Seifert C, Groothuis-Oudshoorn CGM. Comparing three machine learning approaches to design a risk assessment tool for future fractures: predicting a subsequent major osteoporotic fracture in fracture patients with osteopenia and osteoporosis. *Osteoporos Int.* 2021;32:437. <https://doi.org/10.1007/s00198-020-05735-z>.
84. Su Y, Kwok TCY, Cummings SR, Yip BHK, Cawthon PM. Can classification and regression tree analysis help identify clinically meaningful risk groups for hip fracture prediction in older American men (The MrOS Cohort Study)? *JBMR Plus.* 2019;3(10):e10207. <https://doi.org/10.1002/jbm4.10207>.
85. Villamor E, Monserrat C, Del Río L, Romero-Martín JA, Rupérez MJ. Prediction of osteoporotic hip fracture in postmenopausal women through patient-specific FE analyses and machine learning. *Comput Methods Prog Biomed.* 2020;193:105484. <https://doi.org/10.1016/j.cmpb.2020.105484>.
86. Wu Q, Nasoz F, Jung J, Bhattacharai B, Han MV. Machine learning approaches for fracture risk assessment: a comparative analysis of genomic and phenotypic data in 5130 older men. *Calcif Tissue Int.* 2020;107(4):353–61. <https://doi.org/10.1007/s00223-020-00734-y>.
87. Almog YA, Rai A, Zhang P, et al. Deep learning with electronic health records for short-term fracture risk identification: crystal bone algorithm development and validation. *J Med Internet Res.* 2020;22(10):e22550. <https://doi.org/10.2196/22550>.
88. Ma M, Huang DG, Liang X, et al. Integrating transcriptome-wide association study and mRNA expression profiling identifies novel genes associated with bone mineral density. *Osteoporos Int.* 2019;30(7):1521–8. <https://doi.org/10.1007/s00198-019-04958-z>.
89. Ren H, Yu X, Shen G, et al. miRNA-seq analysis of human vertebrae provides insight into the mechanism underlying GIOP. *Bone.* 2019;120:371–86. <https://doi.org/10.1016/j.bone.2018.11.013>.
90. Pillon NJ, Gabriel BM, Dollet L, Smith JAB, Sardón Puig L, Botella J, Bishop DJ, Krook A, Zierath JR. Transcriptomic profiling of skeletal muscle adaptations to exercise and inactivity. *Nat Commun.* 2020;11(1):470. <https://doi.org/10.1038/s41467-019-13869-w>.
91. Li X, Zhang S, Zhang Q, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. *Lancet Oncol.* 2019;20(2):193–201. [https://doi.org/10.1016/s1470-2045\(18\)30762-9](https://doi.org/10.1016/s1470-2045(18)30762-9).
92. Zhao CK, Ren TT, Yin YF, et al. A comparative analysis of two machine learning-based diagnostic patterns with thyroid imaging reporting and data system for thyroid nodules: diagnostic performance and unnecessary biopsy rate. *Thyroid.* 2020;31:470. <https://doi.org/10.1089/thy.2020.0305>.
93. Daniels K, Gummadi S, Zhu Z, Wang S, Patel J, Swendseid B, Lyshchik A, Curry J, Cottrill E, Eisenbrey J. Machine learning by ultrasonography for genetic risk stratification of thyroid nodules. *JAMA Otolaryngol Head Neck Surg.* 2020;146(1):36–41. <https://doi.org/10.1001/jamaoto.2019.3073>.
94. Patel KN, Angell TE, Babiarz J, et al. Performance of a genomic sequencing classifier for the preoperative diagnosis of cytologically indeterminate thyroid nodules. *JAMA Surg.* 2018;153(9):817–24. <https://doi.org/10.1001/jamasurg.2018.1153>.
95. Girolami I, Marletta S, Pantanowitz L, et al. Impact of image analysis and artificial intelligence in thyroid pathology, with particular reference to cytological aspects. *Cytopathology.* 2020;31(5):432–44. <https://doi.org/10.1111/cyt.12828>.
96. Yu J, Deng Y, Liu T, et al. Lymph node metastasis prediction of papillary thyroid carcinoma based on transfer learning radiomics. *Nat Commun.*

- 2020;11(1):4807. <https://doi.org/10.1038/s41467-020-18497-3>.
97. Shao X, Wang B, Mu K, Li L, Li Q, He W, Yao Q, Jia X, Zhang JA. Key gene co-expression modules and functional pathways involved in the pathogenesis of Graves' disease. *Mol Cell Endocrinol.* 2018;474: 252–9. <https://doi.org/10.1016/j.mce.2018.03.015>.
98. Orunesu E, Bagnasco M, Salmaso C, Altrinetti V, Bernasconi D, Del Monte P, Pesce G, Marugo M, Mela GS. Use of an artificial neural network to predict Graves' disease outcome within 2 years of drug withdrawal. *Eur J Clin Investig.* 2004;34(3):210–7. <https://doi.org/10.1111/j.1365-2362.2004.01318.x>.
99. Zaborek NA, Cheng A, Imbus JR, Long KL, Pitt SC, Sippel RS, Schneider DF. The optimal dosing scheme for levothyroxine after thyroidectomy: a comprehensive comparison and evaluation. *Surgery.* 2019;165(1):92–8. <https://doi.org/10.1016/j.surg.2018.04.097>.
100. Wei R, Jiang C, Gao J, et al. Deep-learning approach to automatic identification of facial anomalies in endocrine disorders. *Neuroendocrinology.* 2020;110(5):328–37. <https://doi.org/10.1159/000502211>.
101. Meng T, Guo X, Lian W, Deng K, Gao L, Wang Z, Huang J, Wang X, Long X, Xing B. Identifying facial features and predicting patients of acromegaly using three-dimensional imaging techniques and machine learning. *Front Endocrinol (Lausanne).* 2020;11:492. <https://doi.org/10.3389/fendo.2020.00492>.
102. Kong X, Gong S, Su L, Howard N, Kong Y. Automatic detection of acromegaly from facial photographs using machine learning methods. *EBioMedicine.* 2018;27:94–102. <https://doi.org/10.1016/j.ebiom.2017.12.015>.
103. Saha A, Tso S, Rabski J, Sadeghian A, Cusimano MD. Machine learning applications in imaging analysis for patients with pituitary tumors: a review of the current literature and future directions. *Pituitary.* 2020;23(3):273–93. <https://doi.org/10.1007/s11102-019-01026-x>.
104. Ugga L, Cuocolo R, Solari D, Guadagno E, D'Amico A, Somma T, Cappabianca P, Del Basso de Caro ML, Cavallo LM, Brunetti A. Prediction of high proliferative index in pituitary macroadenomas using MRI-based radiomics and machine learning. *Neuroradiology.* 2019;61(12):1365–73. <https://doi.org/10.1007/s00234-019-02266-1>.
105. Erlic Z, Reel P, Reel S, et al. Targeted metabolomics as a tool in discriminating endocrine from primary hypertension. *J Clin Endocrinol Metab.* 2020;106: 1111. <https://doi.org/10.1210/clinem/dga954>.
106. Chortis V, Bancos I, Nijman T, et al. Urine steroid metabolomics as a novel tool for detection of recurrent adrenocortical carcinoma. *J Clin Endocrinol Metab.* 2020;105(3):e307–18. <https://doi.org/10.1210/clinem/dgz141>.
107. Schweitzer S, Kunz M, Kurlbaum M, Vey J, Kendl S, Deutschbein T, Hahner S, Fassnacht M, Dandekar T, Kroiss M. Plasma steroid metabolome profiling for the diagnosis of adrenocortical carcinoma. *Eur J Endocrinol.* 2019;180(2):117–25. <https://doi.org/10.1530/eje-18-0782>.
108. Wallace PW, Conrad C, Brückmann S, et al. Metabolomics, machine learning and immunohistochemistry to predict succinate dehydrogenase mutational status in phaeochromocytomas and paragangliomas. *J Pathol.* 2020;251(4):378–87. <https://doi.org/10.1002/path.5472>.
109. Burrello J, Burrello A, Pieroni J, et al. Development and validation of prediction models for subtype diagnosis of patients with primary aldosteronism. *J Clin Endocrinol Metab.* 2020;105(10):dga379. <https://doi.org/10.1210/clinend/dga379>.
110. Fan Y, Jiang S, Hua M, Feng S, Feng M, Wang R. Machine learning-based radiomics predicts radiotherapeutic response in patients with acromegaly. *Front Endocrinol (Lausanne).* 2019;10:588. <https://doi.org/10.3389/fendo.2019.00588>.
111. Fan Y, Li Y, Li Y, Feng S, Bao X, Feng M, Wang R. Development and assessment of machine learning algorithms for predicting remission after transsphenoidal surgery among patients with acromegaly. *Endocrine.* 2019. <https://doi.org/10.1007/s12020-019-02121-6>.
112. Hollon TC, Parikh A, Pandian B, Tarpeh J, Orringer DA, Barkan AL, McKean EL, Sullivan SE. A machine learning approach to predict early outcomes after pituitary adenoma surgery. *Neurosurg Focus.* 2018;45(5):E8. <https://doi.org/10.3171/2018.8.Focus18268>.
113. Qiao N, Shen M, He W, et al. Machine learning in predicting early remission in patients after surgical treatment of acromegaly: a multicenter study. *Pituitary.* 2021;24(1):53–61. <https://doi.org/10.1007/s11102-020-01086-4>.
114. Kocak B, Durmaz ES, Kadioglu P, Polat Korkmaz O, Comunoglu N, Tanriover N, Kocer N, Islak C, Kizilkilic O. Predicting response to somatostatin analogues in acromegaly: machine learning-based high-dimensional quantitative texture analysis on T2-weighted MRI. *Eur Radiol.* 2019;29(6):2731–9. <https://doi.org/10.1007/s00330-018-5876-2>.
115. Daye D, Staziaki PV, Furtado VF, et al. CT texture analysis and machine learning improve post-ablation prognostication in patients with adrenal metastases: a proof of concept. *Cardiovasc Intervent Radiol.* 2019;42(12):1771–6. <https://doi.org/10.1007/s00270-019-02336-0>.
116. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25(1):44–56. <https://doi.org/10.1038/s41591-018-0300-7>.
117. Hans D, Shevroja E, Leslie WD. Evolution in fracture risk assessment: artificial versus augmented intelligence. *Osteoporos Int.* 2021;32(2):209–12. <https://doi.org/10.1007/s00198-020-05737-x>.
118. Trischitta V, Copetti M. Moving toward the implementation of precision medicine needs highly discriminatory, validated, inexpensive, and easy-to-use prediction models. *Diabetes Care.* 2020;43(4):701–3. <https://doi.org/10.2337/dc19-0079>.



Artificial Intelligence and Hypertension Management

50

Hiroshi Koshimizu and Yasushi Okuno

Contents

Introduction	690
Artificial Intelligence Approaches for Hypertension Management	691
AI-Surrogate Measurement for BP	691
AI-Factor Analysis for BP Changes	693
AI-Forecasting for Future BP	695
Conclusions	697
References	698

Abstract

The number of hypertensive patients is increasing worldwide. Since high blood pressure is strongly associated with the development of cardiovascular diseases, blood pressure control is essential. More than 90% of hypertensive patients have essential hypertension, which is caused by multiple factors, including lifestyle, physical constitution, and genetics. Blood pressure variability, which is the change in

blood pressure over a certain period, is also associated with cardiovascular diseases. Therefore, regular blood pressure measurements outside the hospital for blood pressure control are required.

The increase in popularity of wearable devices and smartphones has made it easier than ever to gather biometric and environmental information. Blood pressure and lifestyle monitoring using these devices will improve our understanding of the timing of blood pressure rise and fall along with the factors contributing to blood pressure changes. Moreover, the development of novel analysis methods that provide alternatives to conventional statistical methods is expected to improve the accuracy of treatment effect estimations, taking into account individual interpatient differences.

This chapter describes the role of artificial intelligence for blood pressure measurement, factor analysis of blood pressure change, and blood pressure forecasting in personalized

H. Koshimizu

Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto University, Kyoto, Japan

Development Center, Omron Healthcare Co., Ltd., Kyoto, Japan

e-mail: koshimizu.hiroshi.87a@st.kyoto-u.ac.jp

Y. Okuno (✉)

Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto University, Kyoto, Japan

e-mail: okuno.yasushi.4c@kyoto-u.ac.jp

medicine for hypertension management. We summarize the current challenges and future outlooks for using artificial intelligence technologies for hypertension management.

Keywords

Hypertension · Cardiovascular disease · Machine learning · Blood pressure measurement · Blood pressure forecasting · Blood pressure prediction · Personalized medicine · Causal inference · Surrogate model

Introduction

Hypertension is associated with an increased risk of developing cardiovascular diseases (CVD). The number of hypertension patients worldwide has increased from an estimated 594 million in 1975 to 1.13 billion in 2015 [1]. Increasing the number is a significant public health problem. More than 90% of hypertension patients have essential hypertension, which is diagnosed in the absence of a secondary cause. A combination of factors, including lifestyle, physical constitution, and genetics, contribute to the development of essential hypertension [2]. Guidelines for hypertension management in the USA, Europe, and Japan summarize blood pressure (BP) measurement procedures, diagnostic criteria, and treatment strategies [3–5]. However, researchers have described the “hypertension paradox”; over the past decade, the number of hypertension patients has been increasing despite advances in treatment [6]. Moreover, an epidemiological study based on a general Japanese population has reported that 33.1% among hypertensive patients were unaware of their own hypertension [7]; regular BP measurement occasions are necessary. Recently, BP variability, which is short-, medium-, and long-term BP changes, also has been reported as an independent risk factor associated with CVD [8]. Therefore, appropriate treatments based on continuous BP and lifestyle monitoring are necessary to control BP perfectly.

We believe three components are essential for the realization of personalized medicine for

hypertension management: accurately understanding current BP, clarifying appropriate BP control methods, and forecasting future BP. This chapter introduces the following three artificial intelligence (AI) approaches for hypertension management:

1. Surrogate measurement for BP
2. Factor analysis for BP changes
3. Forecasting for future BP

Fig. 1 shows a schematic of these AI. The standard BP measurements for hypertension management are auscultation and oscillometry [3–5]. The disadvantages of these methods include discomfort by the inflated cuff and the need for bulky equipment. Cuffless measurements using AI have been developed as a convenient alternative complement to current BP measurement methods. Easily measurable non-BP indicators were selected as surrogate models for measuring BP. We introduce an AI-surrogate model for measuring BP from pulse waveform by wearable biometric sensors and estimated blood flow using smartphone cameras and discuss the benefits and challenges of these methods.

Next, we review methods for estimating treatment effects and explainable AI (XAI) for hypertension patients as a factor analysis of BP changes. Recent hypertension studies have required analyses that take into account heterogeneous treatment effects (HTE). However, a recent major randomized clinical trial on hypertension management described the limitations of conventional treat-to-target trials [9]. In contrast, AI can estimate HTE among individuals through causal inference using machine learning (ML). XAI is also developed to improve interpretability in AI. In addition to the analysis results, XAI shows the influence of BP-associated factors on an individual’s BP changes. We will describe the application of these AI methods for BP change factor analyses.

Finally, we introduce BP forecasting using AI as a predictive medicine approach for hypertension management. Guidelines for hypertension management recommend measuring BP outside the hospital (out-of-office BP). By forecasting

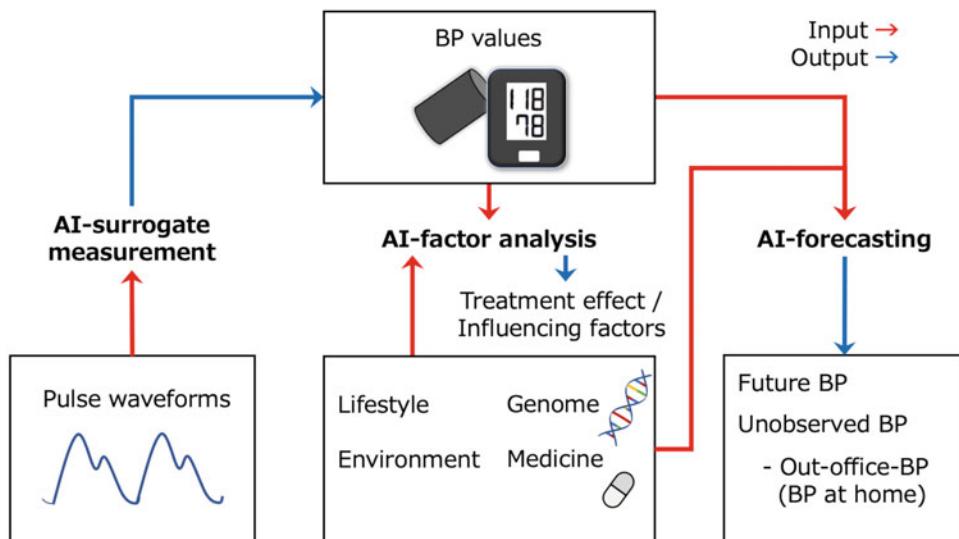


Fig. 1 Scope of artificial intelligence (AI) for hypertension management; AI-surrogate measurement for blood pressure (BP), AI-factor analysis for BP changes, and AI-forecasting for future BP

still unobserved out-of-office BP at the medical examination, physicians can treat hypertension based on the guidelines before patients measure out-of-office BP. In addition, forecasting future BP using time-series data of past BP facilitates preemptive intervention for the patient's future BP. We present the forecasts using cross-sectional data and time-series data.

Artificial Intelligence Approaches for Hypertension Management

AI-Surrogate Measurement for BP

The oscillometric method is used for the standard measurement of out-of-office BP, such as home BP and ambulatory BP monitoring. In the oscillometric method, a cuff wrapped around an upper arm or wrist is inflated with air to obtain the pulse wave amplitude. BP measurement devices calculate systolic and diastolic BP based on changes in the amplitude and the pressure inside the cuff. Discomfort from the inflated cuff and bulky equipment are the main disadvantages of the oscillometric method. BP measurement using pulse transit time (PTT) has been studied in order to achieve cuffless BP measurement [10]. PTT is

the transit time of pulse waves between any two arbitrary arterial sites. PTT is inversely proportional to diastolic BP. The PTT method overcomes the disadvantage of the cuff and provides continuous BP measurement. However, the cardiovascular system provides complex and dynamic feedback in response to BP changes, and its characteristics change over time. For this reason, the PTT method requires calibration for measuring BP.

This section provides surrogate models for measuring BP using an ML as a part of AI, providing a solution to the challenge of oscillometric and PTT methods. PTT methods use waveforms such as electrocardiogram (ECG) and plethysmography (PPG). In this section, we mainly describe PTT methods using the ECG-PPG waveform because consumer devices such as patch-style devices and smartwatches measure these waveforms. The waveforms of ECG and PPG contain hemodynamically relevant indicators such as waveform interval and amplitude in Fig. 2. Surrogate measurements for BP use a combination of signal processing and ML with these raw waveforms and hemodynamically relevant indicators. With improved smartphone cameras and CPUs, convenient BP measurements using smartphones have also been suggested.

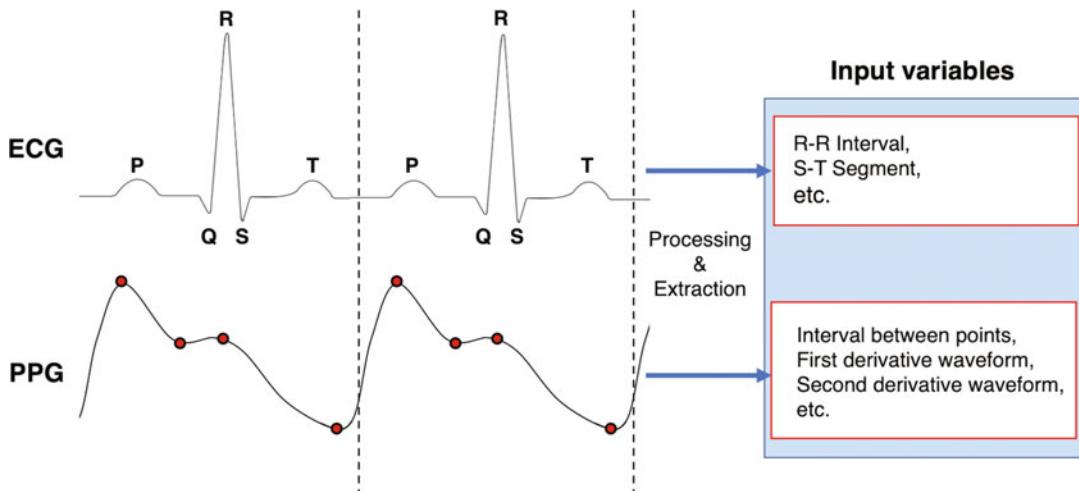


Fig. 2 Input variables for surrogate measurement models for blood pressure using electrocardiography (ECG) and plethysmography (PPG) waveforms

Related studies are described in subsections of this section.

Surrogate BP Measurement Using Wearable Sensors

Calibration is the main challenge in the PTT method using ECG-PPG waveform. Measurement using bidirectional long short-term memory (bidirectional LSTM), a deep neural network (DNN) approaches, was proposed to overcome the challenge [11]. Although DNN approaches were often applied to BP measurement using ECG-PPG waveforms, Su et al. [11] incorporated the residual connection to the DNN for BP surrogate measurement. The residual connection allowed deepening layers of DNNs for high measurement performance without frequent calibration. The proposed method showed that root mean square errors between estimation and reference about systolic and diastolic BP were within 6 mmHg, and the performance was maintained over 6 months.

Additionally, DNNs suitable for time-series data, such as bidirectional LSTM, are often used to measure BP from pulse waveforms [12, 13]. In particular, the attention mechanism, a promising method for improving DNN performance, was powerful for BP measurement [13]. This approach measures BP by inputting the ECG, PPG, and ballistocardiogram (BCG) waveforms without

extracting the amplitude and interval beforehand. Related studies have required much time and are labor intensive for extracting input variables for a surrogate model from waveforms. This measurement method showed the mean absolute error (MAE) and the standard deviation (SD) were 4.06 ± 4.04 and 3.33 ± 3.42 mmHg for systolic and diastolic BP, respectively. In addition, the results in terms of measurement accuracy complied with global standards.

Since some wearable devices can already acquire ECG and PPG waveforms and new sensors will be implemented in the future, we expect to measure BP by incorporating AI into these devices. With the advantage of continuous measurement, wearable devices will conveniently measure BP that is difficult when using cuff-based BP monitors: nocturnal BP during sleep and daytime BP in stressful environments such as the workplace. Therefore, wearable devices will facilitate hypertension detection and the establishment of diagnostic criteria and treatments for BP variability.

Surrogate BP Measurement Using Smartphone Cameras

Conventional BP measurement using ECG-PPG requires a device to capture the waveform data. If common devices such as smartphones could accurately measure BP, hypertension patients who still

have not BP monitor increase the opportunity to aware of BP. For this reason, smartphone-based BP measurements have been developed [14, 15]. Pulse waveforms related to BP are obtained through contact between the smartphone camera or external attachment and the patient's finger. Contactless BP measurement using a smartphone camera has been proposed as a more convenient method [16]. Contactless approaches have used self-captured facial videos; BP is estimated using ML by extracting transdermal blood flow from the videos. The color changes in the facial videos reflect changes in the blood flow under the transdermal conditions; the colors reflect quantities of hemoglobin protein in the blood and melanin pigment in the skin. The proposed method extracted color signals from 17 regions of interest in the videos and estimated BP by applying frequency filtering and ML to the color signals. ML successfully extracted multiple features that effectively estimated BP from those signals. The mean error \pm error SD between systolic and diastolic BP estimation and reference were 0.39 ± 7.30 mmHg and -0.20 ± 6.00 mmHg, respectively. Furthermore, this approach measured BP without requiring calibration or contact with the body. The widespread availability of smartphone-based BP measurement will facilitate early-stage hypertension detection in the future.

Challenges

This section reviewed the available AI approaches for surrogate BP measurement. Although the advantages of these AI methods have been demonstrated, some obstacles remain to be overcome. Validating the accuracy of surrogate BP measurements using AI is challenging because they do not completely comply with existing global standards. Recent studies have discussed the lack of validation for BP measurement accuracy [17, 18]; existing global standards specify measurement procedures and participant attributes in addition to measurement values. Moreover, these standards still do not describe continuous BP monitoring by wearable sensors. Therefore, developing a new validation protocol for continuous BP monitors is necessary to enable global standard compliance of AI-surrogate BP measurement.

AI-Factor Analysis for BP Changes

Hypertension is caused by a combination of factors, including lifestyle, physical constitution, environment, and genetics. Patients with low CVD risk are treated with lifestyle modification, whereas both lifestyle modification and medication are recommended for patients with high CVD risk [3–5]. The Systolic Blood Pressure Intervention Trial (SPRINT) has discussed the target systolic BP for hypertension management [19]. SPRINT research is a landmark study that assessed CVD outcomes on intensive and standard controls for hypertension in patients with high CVD risk. The results show that intensive BP control (systolic BP < 120 mmHg) is more effective than standard BP control (systolic BP < 140 mmHg) for reducing the development of CVD. However, adverse events were reported in the intensive BP control group; thus, physicians are cautious about recommending intensive control measures. Furthermore, the intensity of treatment (e.g., dosage) differs among individual patients, and analyses that take into account heterogeneous treatment effects (HTE) are required. Conventional randomized clinical trial approaches cannot assign enough participants into subgroups for estimating HTE. For this background, a previous study has suggested a new HTE estimation [9].

This section reviews causal inference using ML and XAI approaches for hypertension management. Causal inference using ML has been developed for the HTE estimation not only to hypertension management but to various fields [20]. Additionally, the XAI approach unravels the black box reasoning, which facilitates analysis of individual patient's hypertension factors. In subsections, we present previous studies about factor analyses for BP changes.

Causal Inference Using ML for Hypertension

Causal forests are random forest (RF)-based methods for HTE estimation [20]. RFs are decision-tree algorithms that cluster hierarchically like a tree diagram. Samples with similar attributes (e.g., patients) are assigned to nodes in the tree diagram. Causal forests use nodes of RF to assess the treatment effects among samples within

the same node. This method is used to tackle the challenge of subgroup analysis in randomized clinical trial datasets. Similar RF- or decision-tree-based methods have been developed for HTE estimation [21]. Scarpa et al. [22] identified a group of adverse events in the SPRINT using HTE estimation results from an RF-based algorithm. In the intensive BP control group, significantly higher incidences of CVD were detected in patients with both systolic BP ≥ 144 mmHg who also smoked.

Künzel et al. [23] developed X-learner to estimate HTE on an individual basis rather than a group basis, using a method that does not use causal forests. X-learner uses some regressions to calculate the individual treatment effect on a counterfactual framework. These regressions took into account the effect of nonlinear patient backgrounds using ML; X-learner was used for secondary analysis to the SPRINT datasets [24]. Duan et al. [24] compared the performance of treatment effect estimation using logistic regression and X-learner. The results showed that X-learner accurately predicted reductions in CVD risk within 3 years of treatment. However, logistic regression showed that the high baseline risk was proportional to a high treatment effect, but X-learner showed that it was not. Thus, X-learner-estimated HTE was not associated with baseline CVD risk and was not considered by linear models.

Causal inference using ML is rapidly evolving within the computer science field and can also be applied in various fields outside of medicine. This method will be the key to developing personalized medicine for hypertension management.

Explainable AI for Hypertension

Causal inference using ML focuses on estimating HTE of the intervention (e.g., lifestyle modification and medications). ML also accurately predicts incident hypertension and BP values by training big data related to hypertension. However, ML is a black box approach; big data inputs into nonlinear models make interpreting outputs difficult. XAI has been developed for understanding the black box.

Elshawi et al. compared the interpretability of XAI against the hypertension prediction model by ML [25]. The dataset consisted of 43 variables related to hypertension from 23,095 participants. Seven methods for interpreting the ML prediction model were applied, and two among seven methods could effectively evaluate individual hypertension prediction variables: Local Interpretable Model-agnostic Explanations (LIME) [26] and SHapley Additive exPlanations (SHAP) [27]. The experiment results showed that LIME is unstable because it produced different interpretations for two individuals with similar backgrounds. SHAP was a more suitable method because it enabled comparisons between individuals on an evaluation axis using SHAP value for each variable.

SHAP also have been used to evaluate the effect of hypertension management apps [28]. App-based hypertension treatment is expected to affect treatment continuity and BP reduction. The study predicted whether app-based hypertension treatment would be continued for 10 weeks and identified variables that contribute to treatment continuation. The dataset consisted of 427 participants on days one, three, and seven of the study. The dataset included 19 variables, including smartphone OS and email domain as social background variables, the number of plant-based meals reflecting engagement with the treatment, and weight and BP as biometrics. This study built a prediction model for hypertension management continuity using RF. SHAP was used to interpret the individual effect of the model. The results based on SHAP suggested that continuous weight measurement, exercise reports, smartphone OS, and email domain of study participant were related to whether the app-based hypertension treatment would be continued for 10 weeks. SHAP provided new insights from the ML model, which could not be interpreted using black box approaches.

Challenges

Causal inference using AI, and XAI was demonstrated to be a powerful tool in hypertension research. However, many clinical databases on hypertension to which the described AIs could

be applied are not publicly available. Clinical databases are collected on human subjects; thus, the cost for collecting and managing data is high, and data protection and privacy regulations also are strict. To address this issue, a clinical dataset generation approach using DNNs has been proposed [29]. This approach can generate clinical data while protecting patient privacy. If datasets similar to real-world clinical data can be created and made publicly available, AI-factor analyses will facilitate the establishment of BP management evidence for personalized optimal treatment.

AI-Forecasting for Future BP

Hypertension is diagnosed in patient's office/hospital systolic BP of ≥ 140 mmHg and/or diastolic BP of ≥ 90 mmHg, or out-of-office systolic BP of ≥ 135 mmHg and/or diastolic BP of ≥ 85 mmHg [3–5]. If physicians could accurately forecast future or unobserved BP, BP could be managed based on the forecast. Previous studies have proposed out-of-office BP predictions using age, height, weight, lifestyle, medical history, and office/hospital BP measurements. A recent study compared new and existing strategies for predicting out-of-office BP in the context of diagnostic costs related to hypertension [30]. Existing diagnostic strategies use office/hospital BP for an initial decision. If the office/hospital BP of a hypertension-suspected patient exceeds diagnostic criteria, the patient has to measure out-office BP for a certain period. Physicians then diagnose whether hypertension based on the out-office BP requires the patient to continue measuring out-office BP for BP management as a final decision. In contrast, the new diagnostic strategy uses predicted difference between office/hospital BP and out-of-office BP based on individual's characteristics; thus, physicians diagnose hypertension by using prediction value instead of out-office BP for a certain period as a final decision. The study suggested that the new diagnostic strategy for hypertension management is more cost-effective than existing diagnostic strategies.

In hypertension research, BP variability has been identified as an independent risk factor for

CVD [8]. BP variability is defined as the SD and coefficient of BP variation over a certain period. However, the periods vary from beat-to-beat, day-to-day, and office/hospital visit-to-visit. Therefore, treatments and criteria for BP variability have not been established. BP variability is a potential indicator of future BP that can guide management recommendations. Variables of BP time-series data are related to past BP change patterns and are useful for forecasting future BP as shown in Fig. 3. The following subsections describe BP/hypertension forecasting approaches and challenges.

BP Forecasting with Cross-Sectional Data

We can describe future BP or hypertension forecasts using BP-related cross-sectional data. Essential hypertension, which is caused by a combination of factors, accounts for approximately 90% of all hypertension cases [2]. The influence of nongenetic factors can be significant, and even patients with hypertension-related genetic factors can achieve lower BP via lifestyle modification [31]. Therefore, high-performance hypertension forecasting requires datasets including congenital genetic factors as well as acquired lifestyle and physical characteristics such as BMI, age, and sex.

Pei et al. [32] proposed a method for incident hypertension prediction using genetic and environmental factors. Twelve single-nucleotide polymorphism markers were included as genetic factors, and height, weight, gender, age, occupation, smoking status, drinking situation, family history of hypertension, and family history of coronary heart disease were included as environmental factors. Three hypertension prediction models were prepared using an ML: one using only genetic factors, one using only environmental factors, and one using genetic and environmental factors. The dataset consisted of 559 patients with hypertension and 641 healthy individuals. The best prediction performance was obtained using the model that included genetic and environmental factors. The sensitivity and specificity of the model were 63.3% and 86.7%, respectively. The sensitivity and specificity of the genetic-factor-only and environmental-factor-only

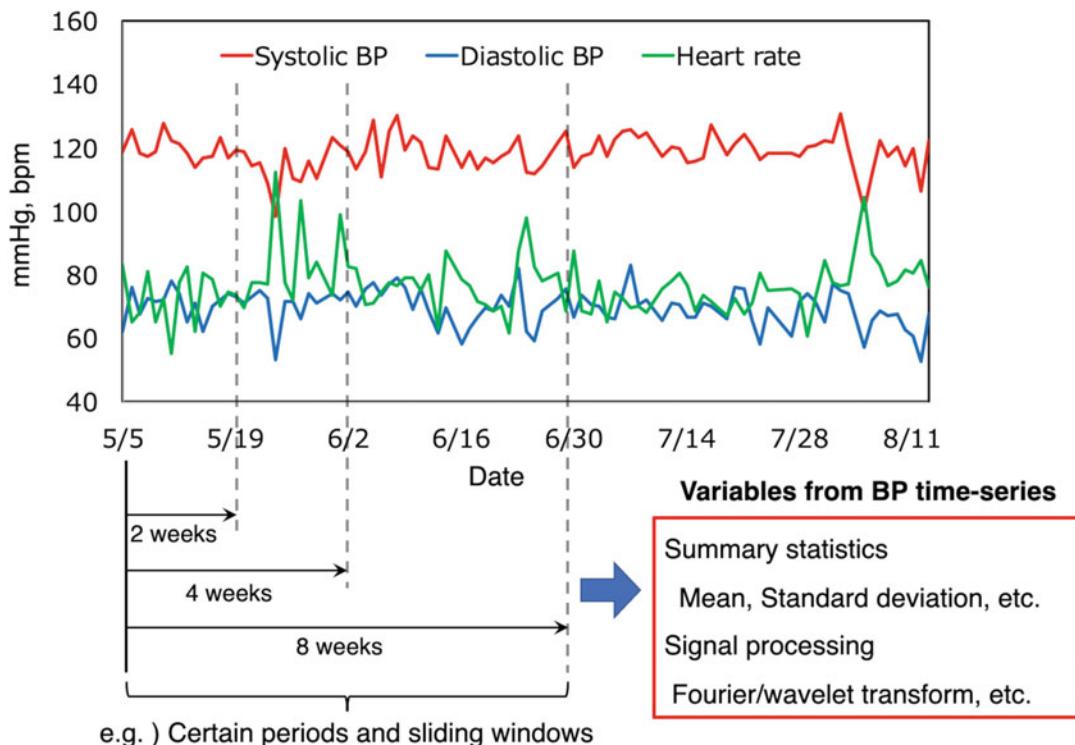


Fig. 3 Example variables of blood pressure (BP) change patterns for future BP forecasting

models were 65.5% and 45.8% and 59.3% and 85.1%, respectively. These results suggested that environmental factors were better hypertension predictors than genetic factors when using this ML prediction approach.

Yang et al. [33] reported incident hypertension prediction within 1–3 years based on lifestyle and physical constitution. The study included 34,719 participants in the US working population, aged 18–54 years. Hypertension in the working population has a greater economic impact due to health care expenditure and hypertension-related productivity reduction. A multistate Markov model was implemented to predict the underlying transition patterns among BP states; namely, normal ($SBP < 120$ mmHg and/or $DBP < 80$ mmHg), elevated ($SBP 120\text{--}129$ mmHg and/or $DBP < 80$ mmHg), and hypertensive ($SBP \geq 130$ mmHg and/or $DBP \geq 80$ mmHg or taking antihypertensive medicine) states, as defined by the American College of Cardiology/American Heart Association (2017

ACC/AHA) guideline [3]. The Markov model is a classical stochastic process in which the state at any time depends only on the previous state. The Markov model predictions were consistent with common hypertension findings, showing that obese and male patients were more likely to develop hypertension. Yang et al. [33] also developed a web application to simulate the probability of developing hypertension with weight loss. Future work will aim to validate the effectiveness of the application as a decision support tool for health care professionals.

Ye et al. [34] also developed an ML-based incident hypertension prediction model. The 1-year incident hypertension risk model was constructed using statewide electronic health records from a retrospective cohort of 823,627 individuals. The model performance was then tested on a prospective cohort of 680,810 individuals. XGBoost [35] was selected to achieve high-performance hypertension prediction. XGBoost is

a decision-tree-based ML approach, known as a high-performance algorithm for regression and classification tasks. The model obtained area under curves (AUCs) of 0.917 and 0.870 in the retrospective and prospective cohorts, respectively. The incident hypertension prediction model has already been deployed in some states; thus, future studies will investigate the health care cost reduction and patient outcome improvement achieved using this model.

BP Forecasting with Time-Series Data

BP can be forecasted using time-series data such as visit-to-visit and day-to-day BP measurements. A previous study has reported daily BP prediction by the statistical method as complement method unobserved BP of time-series data [36]. This subsection introduces BP forecasting using ML approaches.

Lacson et al. [37] analyzed the associations between changes in the SPRINT BP time-series data, including BP variability and CVD development. In addition to the summary statistics, time-series data included various variables such as periodicities and amplitudes for certain periods. The model extracted individual BP characteristics from BP time-series data by wavelet transform signal processing. CVD onset was predicted from the extracted BP data using RF; the CVD prediction model obtained an AUC of 0.71, BP time-series data after wavelet transformation, which is a feature of time-series changes in BP, were effective for CVD prediction. This result suggests that BP time-series data can improve the performance of BP and CVD-related prediction. Therefore, BP forecasting, which aims to prevent CVD development, should incorporate time-series BP data.

Li et al. [38] developed an approach to predict future BP from past BP time-series data using DNNs. The prediction model is implemented LSTM [39], a DNN for time-series data, to predict the monthly mean BP 1–3 months in the future. BP and BMI were time-series data inputs, and other BP-related factors were included as snapshot data. The MAE between the reference and predicted BP was within 7 mmHg and 5 mmHg

for systolic and diastolic BP, respectively. If BP is predicted sequentially, as in this approach, it will complement BP when hypertension patients did not take measurements. Furthermore, daily BP forecasts with high performance will enable physicians to understand BP variability beforehand. Hypertension management based on BP forecasts using time-series data will provide novel patient stratification using BP change patterns for personalized medicine.

Challenges

This section reviewed AI applications for future and unobserved BP forecasts. High-performance BP forecasts are useful for estimating future health care costs. However, the clinical significance of BP forecasts has not been validated. BP and BP-related lifestyle and environment data can be continuously acquired from wearable devices and smartphones; therefore, it is essential to validate BP forecasting in the real world as well as the laboratory. Moreover, daily and weekly BP time-series databases are rarely available. From this viewpoint, as described in section “[AI-Factor Analysis for BP Changes](#),” database generation [29] is required to accelerate research regarding BP forecasting using time-series data.

Conclusions

This chapter discussed the current research on AI for surrogate measurement for BP, factor analysis of BP changes, and BP forecasting for hypertension management (Table 1). Surrogate BP measurements enable continuous and convenient BP measurement to recognize when patients experience rising BP. Factor analyses of BP changes can facilitate the realization of personalized medicine for hypertension management by estimating individual treatment effects. BP forecasting will be advanced by the increasing availability of BP and BP-related data acquired from wearable devices. In the future, these AI approaches for hypertension management will contribute to reducing hypertension incidence and associated health care costs.

Table 1 Artificial intelligence (AI) approaches for hypertension management

AI approaches	Categories	Authors	Year	Contents
AI-surrogate measurement for BP	Measurements using wearable sensors	Su P et al. [11]	2018	Systolic and diastolic BP estimation using ECG-PPG waveforms.
		Franco G et al. [12]	2019	Systolic and diastolic BP (continuous BP waveforms) estimation using ECG-PPG waveforms.
		Eom H et al. [13]	2020	Systolic and diastolic BP estimation using ECG, PPG, and BCG waveforms.
	Measurement using smartphone cameras	Luo H et al. [16]	2019	Systolic and diastolic BP, pulse pressure estimation using facial videos.
AI-factor analysis for BP changes	Causal inference	Powers S et al. [21]	2018	Suggestion of HTE estimation methods on SPRINT study.
		Scarpa J et al. [22]	2019	Identifying using ML of attributes in adverse event group on SPRINT study.
		Duan T et al. [24]	2019	HTE estimation in an individual, comparison of causal Inferences by logistic regression and ML.
	Explainable AI	Elshawi R et al. [25]	2019	Comparisons of interpretability by explainable AI.
		Guthrie NL et al. [28]	2019	Exploration of the relationship between long-term engagement and mobile app intervention using AI.
AI-forecasting for future BP	Forecasting with cross-sectional data	Pei Z et al. [32]	2018	Incident hypertension prediction from genetic and environmental factors using ML.
		Yang J et al. [33]	2020	Incident hypertension prediction based on stochastic process and its application to physician support tool.
		Ye C et al. [34]	2018	Incident hypertension prediction using large-scale electronic health records.
	Forecasting with time-series data	Lacson RC et al. [37]	2019	CVD prediction using BP time-series data including BP variability.
		Li X et al. [38]	2017	Future BP forecasting using past BP and BMI time-series data.

BP, blood pressure; ECG, electrocardiography; PPG, plethysmography; BCG, ballistocardiogram; HTE, heterogeneous treatment effects; SPRINT, Systolic Blood Pressure Intervention Trial; ML, machine learning; CVD, cardiovascular disease

References

- Zhou B, Bentham J, Di Cesare M, Bixby H, Danaei G, Cowan MJ, et al. Worldwide trends in blood pressure from 1975 to 2015: a pooled analysis of 1479 population-based measurement studies with 19·1 million participants. Lancet. 2017;389(10064):37–55. [https://doi.org/10.1016/S0140-6736\(16\)31919-5](https://doi.org/10.1016/S0140-6736(16)31919-5).
- Carretero OA, Oparil S. Essential hypertension. Part I: Definition and etiology. Vol. 101, Circulation. Lippincott Williams and Wilkins; 2000. p. 329–35. <https://doi.org/10.1161/01.CIR.101.3.329>.
- Whelton PK, Carey RM, Aronow WS, Casey DE, Collins KJ, Dennison Himmelfarb C, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Hypertension. 2018;71(6):E13–115. <https://doi.org/10.1161/HYP.000000000000065>.
- Williams B, Mancia G, Spiering W, Agabiti Rosei E, Azizi M, Burnier M, et al. 2018 ESC/ESH guidelines for the management of arterial hypertension. Eur Heart J. 2018;39(33):3021–104. <https://doi.org/10.1093/eurheartj/ehy339>.
- Umemura S, Arima H, Arima S, Asayama K, Dohi Y, Hirooka Y, et al. The Japanese Society of hypertension guidelines for the management of hypertension (JSH 2019). Hypertens Res. 2019;42(9):1235–481. <https://doi.org/10.1038/s41440-019-0284-9>.
- Chobanian AV. The hypertension paradox – more uncontrolled disease despite improved therapy. N

- Engl J Med. 2009;361(9):878–87. <https://doi.org/10.1056/NEJMsa0903829>.
7. Satoh A, Arima H, Ohkubo T, Nishi N, Okuda N, Ae R, et al. Associations of socioeconomic status with prevalence, awareness, treatment, and control of hypertension in a general Japanese population: NIPPON DATA 2010. *J Hypertens.* 2017;35(2):401–8.
8. Stevens SL, Wood S, Koshiaris C, Law K, Glasziou P, Stevens RJ, et al. Blood pressure variability and cardiovascular disease: systematic review and meta-analysis. *BMJ.* 2016;354:i4098. <https://doi.org/10.1136/bmj.i4098>.
9. Basu S, Sussman JB, Hayward RA. Detecting heterogeneous treatment effects to guide personalized blood pressure treatment. *Ann Intern Med.* 2017;166(5):354. <https://doi.org/10.7326/M16-1756>.
10. Mukkamala R, Hahn J-O, Inan OT, Mestha LK, Kim C-S, Toreyin H, et al. Toward ubiquitous blood pressure monitoring via pulse transit time: theory and practice. *IEEE Trans Biomed Eng.* 2015;62(8):1879–901. <https://doi.org/10.1109/TBME.2015.2441951>.
11. Su P, Ding X-R, Zhang Y-T, Liu J, Miao F, Zhao N. Long-term blood pressure prediction with deep recurrent neural networks. In: 2018 IEEE EMBS international conference on biomedical & health informatics (BHI). IEEE; 2018. p. 323–8. <https://doi.org/10.1109/BHI.2018.8333434>.
12. Franco G, Cerina L, Gallicchio C, Micheli A, Santambrogio MD. Continuous blood pressure estimation through optimized echo state networks. In: International conference on artificial neural networks. Springer International Publishing; 2019. p. 48–61. https://doi.org/10.1007/978-3-030-30493-5_5.
13. Eom H, Lee D, Han S, Hariyani YS, Lim Y, Sohn I, et al. End-to-end deep learning architecture for continuous blood pressure estimation using attention mechanism. *Sensors.* 2020;20(8):2338. <https://doi.org/10.3390/s20082338>.
14. Chandrasekhar A, Kim C-S, Naji M, Natarajan K, Hahn J-O, Mukkamala R. Smartphone-based blood pressure monitoring via the oscillometric finger-pressing method. *Sci Transl Med.* 2018;10(431):eaap8674. <https://doi.org/10.1126/scitranslmed.aap8674>.
15. Schoettker P, Degott J, Hofmann G, Proen  a M, Bonnier G, Lemkadem A, et al. Blood pressure measurements with the OptiBP smartphone app validated against reference auscultatory measurements. *Sci Rep.* 2020;10(1):17827. <https://doi.org/10.1038/s41598-020-74955-4>.
16. Luo H, Yang D, Barszczuk A, Vempala N, Wei J, Wu SJ, et al. Smartphone-based blood pressure measurement using transdermal optical imaging technology. *Circ Cardiovasc Imaging.* 2019;12(8):8857. <https://doi.org/10.1161/CIRCIMAGING.119.008857>.
17. Stergiou GS, Alpert BS, Mieke S, Wang J, O'Brien E. Validation protocols for blood pressure measuring devices in the 21st century. *J Clin Hypertens.* 2018;20(7):1096–9. <https://doi.org/10.1111/jch.13294>.
18. D  rr M, Weber S, Birkemeyer R, Leonardi L, Winterhalder C, Raichle CJ, et al. iPhone app compared with standard blood pressure measurement –the iPARR trial. *Am Heart J.* 2021;233:102–8. <https://doi.org/10.1016/j.ahj.2020.12.003>.
19. SPRINT Research Group. A randomized trial of intensive versus standard blood-pressure control. *N Engl J Med.* 2015;373(22):2103–16. <https://doi.org/10.1056/NEJMoa1511939>.
20. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. *J Am Stat Assoc.* 2015;110(523):1228–42. <https://doi.org/10.1080/01621459.2017.1319839>.
21. Powers S, Qian J, Jung K, Schuler A, Shah NH, Hastie T, et al. Some methods for heterogeneous treatment effect estimation in high dimensions. *Stat Med.* 2018;37(11):1767–87. <https://doi.org/10.1002/sim.7623>.
22. Scarpa J, Bruzelius E, Doupe P, Le M, Faghmous J, Baum A. Assessment of risk of harm associated with intensive blood pressure management among patients with hypertension who smoke. *JAMA Netw Open.* 2019;2(3):e190005. <https://doi.org/10.1001/jamanetworkopen.2019.0005>.
23. K  nzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc Natl Acad Sci.* 2019;116(10):4156–65. <https://doi.org/10.1073/pnas.1804597116>.
24. Duan T, Rajpurkar P, Laird D, Ng AY, Basu S. Clinical value of predicting individual treatment effects for intensive blood pressure therapy. *Circ Cardiovasc Qual Outcomes.* 2019;12(3):e005010. <https://doi.org/10.1161/CIRCOUTCOMES.118.005010>.
25. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. *BMC Med Inform Decis Mak.* 2019;19(1):146. <https://doi.org/10.1186/s12911-019-0874-0>.
26. Ribeiro MT, Singh S, Guestrin C. “Why should I trust you?” explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM; 2016. p. 1135–44. <https://doi.org/10.1145/2939672.2939778>.
27. Lundberg S, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst.* 2017;2017:4766–75.
28. Guthrie NL, Berman MA, Edwards KL, Appelbaum KJ, Dey S, Carpenter J, et al. Achieving rapid blood pressure control with digital therapeutics: retrospective cohort and machine learning study. *JMIR Cardio.* 2019;3(1):e13030. <https://doi.org/10.2196/13030>.
29. Beaulieu-Jones BK, Wu ZS, Williams C, Lee R, Bhavnani SP, Byrd JB, et al. Privacy-preserving generative deep neural networks support clinical data sharing. *Circ Cardiovasc Qual Outcomes.* 2019;12(7):5122. <https://doi.org/10.1161/CIRCOUTCOMES.118.005122>.
30. Monahan M, Jowett S, Lovibond K, Gill P, Godwin M, Greenfield S, et al. Predicting out-of-office blood

- pressure in the clinic for the diagnosis of hypertension in primary care. *Hypertension*. 2018;71(2):250–61. <https://doi.org/10.1161/HYPERTENSIONAHA.117.10244>.
31. Pazoki R, Dehghan A, Evangelou E, Warren H, Gao H, Caulfield M, et al. Genetic predisposition to high blood pressure and lifestyle factors: associations with midlife blood pressure levels and cardiovascular events. *Circulation*. 2018;137(7):653–61. <https://doi.org/10.1161/CIRCULATIONAHA.117.030898>.
32. Pei Z, Liu J, Liu M, Zhou W, Yan P, Wen S, et al. Risk-predicting model for incident of essential hypertension based on environmental and genetic factors with support vector machine. *Interdiscip Sci Comput Life Sci*. 2018;10(1):126–30. <https://doi.org/10.1007/s12539-017-0271-2>.
33. Yang J, Liu F, Wang B, Chen C, Church T, Dukes L, et al. Blood pressure states transition inference based on multi-state Markov model. *IEEE J Biomed Heal Informatics*. 2020;25(1):237–46. <https://doi.org/10.1109/JBHI.2020.3006217>.
34. Ye C, Fu T, Hao S, Zhang Y, Wang O, Jin B, et al. Prediction of incident hypertension within the next year: prospective study using statewide electronic health records and machine learning. *J Med Internet Res*. 2018;20(1):e22. <https://doi.org/10.2196/jmir.9268>.
35. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM; 2016. p. 785–94. <https://doi.org/10.1145/2939672.2939785>.
36. Campos LF, Glickman ME, Hunter KB. Measuring effects of medication adherence on time-varying health outcomes using Bayesian dynamic linear models. *Biostatistics*. 2019;1–22. <https://doi.org/10.1093/biostatistics/kxz059>.
37. Lacson RC, Baker B, Suresh H, Andriole K, Szolovits P, Lacson E. Use of machine-learning algorithms to determine features of systolic blood pressure variability that predict poor outcomes in hypertensive patients. *Clin Kidney J*. 2019;12(2):206–12. <https://doi.org/10.1093/ckj/sfy049>.
38. Li X, Wu S, Wang L. Blood pressure prediction via recurrent models with contextual layer. In: Proceedings of the 26th international conference on world wide web. Republic and Canton of Geneva: International World Wide Web Conferences Steering Committee; 2017. p. 685–93. <https://doi.org/10.1145/3038912.3052604>.
39. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9(8):1735–80. <https://doi.org/10.1162/neco.1997.9.8.1735>.



Aim and Diabetes

51

Josep Vehi, Omer Mujahid, and Ivan Contreras

Contents

Introduction	702
Problems Faced by People with Diabetes	703
The AI Approach	703
Expert Systems Versus Decision Support Systems	704
Prevention and Prognosis	705
Diagnosis of Diabetes	705
Diabetes Management	705
Lifestyle Interventions/Behavioral Change	706
Risk Stratification	707
Complications and Comorbidities	707
Discussion	707
Conclusions	708
References	708

Abstract

Artificial Intelligence is set to revolutionize diabetes health care. Diabetes is the increase in blood glucose levels above normal range. It is tied to the human body's inability to produce insulin or utilize it. Diabetes has taken form of a global pandemic with over 463 million people suffering from it, worldwide. The current health care system is finding it hard to keep up with the demands of diabetes with the existing structure proving to be both expensive and ineffective. In artificial intelligence, we have a tool that has the capability of transforming the current system for good. The shift has already started to take place with many scientific studies focused on artificial intelligence

J. Vehi (✉)

Modelling, Identification and Control Engineering Laboratory (MICELab), Institut d'Informatica i Aplicacions, Universitat de Girona, Girona, Spain

Centro de Investigación Biomédica en Red de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM), Barcelona, Spain
e-mail: josep.veh@udg.edu

O. Mujahid · I. Contreras

Modelling, Identification and Control Engineering Laboratory (MICELab), Institut d'Informatica i Aplicacions, Universitat de Girona, Girona, Spain
e-mail: omer.mujahid@udg.edu; ivancontreras@udg.edu

for diabetes management are being realized into commercialized products that are replacing existing methods of diabetes treatment and management. Artificial intelligence finds its use in various aspects of diabetes health care that benefits both the patients and diabetes health care professionals. This chapter dwells in on all the problems faced by diabetics and how artificial intelligence can come up with solutions for these problems.

Keywords

Artificial intelligence · Diabetes · Machine learning · Decision support systems · Expert systems

Introduction

Artificial intelligence (AI) is set to bring about a radical alteration in medicine and health care [1]. In diabetes, in particular, AI promises a lot of good things. Diabetes is the increase in blood glucose (BG) levels above a critical limit in the human body. The human body contains certain regulatory mechanisms that keep the body's BG levels in an optimal range. The healthy range of BG is 70 mg/dL to 140 mg/dL. It is the function of hormones like insulin and glucagon to keep the BG levels regulated with insulin keeping the higher levels in check by absorbing glucose and glucagon taking care of low BG levels by converting stored glycogen into glucose when needed. It is when the BG levels surpass the higher limit of 140 mg/dL that diabetes is diagnosed. Diabetes may be categorized in multiple types with type 1 diabetes (T1D), type 2 diabetes (T2D), and gestational diabetes being three of its main classifications [2].

T1D is an autoimmune disease in which the pancreas produces little or no insulin [3]. It is caused when the insulin producing beta cells are destroyed by the immune response of human body. On the other hand, T2D is caused when the body cells do not respond to insulin and hence the glucose level keeps on increasing in the human blood stream [4]. The exact reason of why T1D is caused is unknown and is majorly

believed to be genetic. While the cause of T2D is also been connected to genetic factors, it is also believed that environmental factors and obesity are contributors to its occurrence as well. On the other hand, gestational diabetes is the increase in BG levels of women during pregnancy [5]. The reason for this is the production of hormones by placenta for sustaining pregnancy that causes cell resistance to insulin. In most cases, gestational diabetes is cured after pregnancy.

It is understood that diabetes is a disease that cannot be cured completely. It can, however, be controlled by medications and a change in lifestyle. This is the reason that the term diabetes management has been coined. Diabetes management refers to a group of different measures taken to regulate BG and increase the time of a person's BG profile in normoglycemic range. Over the years, multiple methods and tools have been used for the purpose of diabetes management. Technological advances like Internet, wireless communication, and high-end sensors have aided to the goal of a highly efficient diabetes management system that is capable of keeping the BG profile of a diabetic patient in the normoglycemic range all the time. The field of AI holds promise in coming up with solutions to the problems faced by patients suffering from diabetes and the diabetes health care professionals.

AI involves methods and techniques that enable a computer program to imitate human ways in intelligently figuring out the outcomes of various processes. With the advent of sensors such as the continuous glucose monitor (CGM), the acquisition of blood glucose (BG) data became possible and with it the doors of artificial intelligence in diabetes opened. In the past, majority of AI in health care was dominated by methods like case-based reasoning (CBR) and rule-based reasoning (RBR) [6]. However, it was with advances in machine learning (ML) and artificial neural networks (ANN) that revolutionized AI. AI owe most this success to the availability of large amount of data. From expert systems (ES) to decision support systems (DSS) to prediction and diagnostic tools, AI provides solutions to transform almost every aspect of diabetes health care.

This chapter provides an overview of the latest advances in AI in diabetes. The discussion commences with an account of the problems faced by

diabetic patients followed by the solutions provided by AI to these problems. The solutions are then subcategorized in different areas of diabetes health care that can be targeted with the help of AI. Rest of the chapter is distributed as: section “[Problems Faced by People with Diabetes](#)” I explains the problems faced by diabetic patients. Section “[The AI Approach](#)” explores the solutions in AI for the problems discussed in section “[Problems Faced by People with Diabetes](#).“ Section “[Discussion](#)” provides a discussion of the entire chapter and a conclusion to the chapter is given in section “[Conclusions](#).“

Problems Faced by People with Diabetes

Diabetes is a world pandemic with over 463 million people suffering from it, worldwide. This figure is expected to rise to 700 million by 2045. In 2019 alone diabetes was cause to 4.2 million deaths [7]. What is even more alarming that every 1 in 2 people is undiagnosed. Moreover, people suffering from diabetes experience various adversities in the form of physical and mental traumas that considerably reduce the quality of their lives. It is, thus, a matter of critical importance that measures be taken to control the devastating effect of this disease and to improve the life quality of people who are suffering from it.

The occurrence of diabetes in a human may lead to a multitude of different complications. These complications are often referred to as long-term comorbidities. When a patient suffers from diabetes for a longer duration and takes no precaution for BG regulation, these complications are very likely to arise. Cardiovascular diseases, neuropathy, retinopathy, nephropathy, foot damage, skin conditions, hearing impairments, Alzheimer’s disease, and depression can all be the comorbidities arising because of diabetes mellitus. In T1Ds the occurrence of hypoglycemia is a major complication. It occurs when the BG levels go below the critical levels. The happening of hypoglycemia may cause loss of consciousness, confusion, trouble talking, seizures, and in extreme cases death.

Although, T1D is genetic and cannot be prevented, there is a chance T2D may be prevented.

As already discussed, the cause of diabetes is often unknown in all cases. In T2D, however, prediabetes is an indicator of a future occurrence of diabetes. Prediabetes is the rise in BG levels above normal levels and can serve as a warning to the patient prior to the diagnosis of diabetes. Gestational diabetes can also lead to T2D in some cases. The diagnosis of diabetes itself is a problem faced by many patients. This is confirmed by the fact that almost half of all the diabetes cases around the world are diagnosed late or not diagnosed at all.

Keeping the BG levels in normoglycemic range is one of the major challenges in diabetes management. For T1Ds, management of the insulin intake regime is necessary. The amount, time, and intensity of basal and bolus insulin need to be specified according to the dynamics of a patient’s BG profile. The management of insulin is also important for T2D patients on insulin. In T2D, the patients need to have a thorough idea of their diet and other lifestyle choices that may lead to complications. Diabetic patients are prone to mental stress. In case of T1D the unpredictability of events like hypoglycemia can add immensely to the fear and stress of a patient. This can reduce the life quality of a patient and may trigger other mental issues such as depression. Moreover, stress is also understood to be one of the factors that increases the risk of T2D.

With the increase in global population, our health care system finds it hard to keep up with the health care demands of the increasing number of diabetes patients. The existing infrastructure does not allow for a diabetic patient to enjoy the attention they deserve. Limited diagnostic facilities and a shortage of medications worldwide have lead us to a point where the demand for a radical shift in the current system seems obligatory. AI can provide us with the platform for the shift in diabetes health care that is utterly necessary.

The AI Approach

The term AI has been around for a good five decades now. It was first coined by the ground breaking science article of Alan Turing entitled “Computational Machinery and Intelligence.” There is no single definition of AI that is

universally accepted. However, the authors of “Artificial Intelligence: A Modern Approach,” a landmark book in the field of AI, defines AI as “the study of agents that receive precepts from the environment and perform actions.” With advances in ML and deep learning (DL) in the past decade, this field of computer science has taken shape of a revolution. Even though AI finds its use in a large group of applications, health care in general and diabetes health care in particular can benefit from it greatly. The problems discussed in the last section can all be addressed with AI-based techniques. Some of the major AI-based solutions in diabetes health care are stated next.

Expert Systems Versus Decision Support Systems

As shown in Fig. 1, almost all of the AI-based tools designed for diabetes health care will eventually find their use in either ES or DSS or both. Diabetes ES are systems that are designed with a focus on mimicking the real-life diabetes expert.

These systems determine the condition of a diabetes patient by analyzing different parameters related to the physiological condition of a patient. These parameters are often obtained in form of laboratory tests, electronic health records (EHRs), patient secure messages, or questionnaires. The ES of past were mainly based on RBR or CBR and followed an if-else conditional structure. The CBR-based ES look at similar cases from the past and determine the outcome of a specific case, whereas RBR-based ES determine the outcome of a case based on certain rules. These rules are similar to the ones used by clinicians while analyzing the condition of a patient by questioning them. With advances in ML and DL, many ES now utilize algorithms that learn from data.

DSS on the other hand are software platforms used to help the diabetic patients and the clinicians with their decisions regarding a certain scenario by giving out suggestions. The suggestions DSS give a diabetic patient or clinician are often based on outcomes obtained from AI-based models. These suggestions when adopted in the course of action are supposed to improve the outcome of a

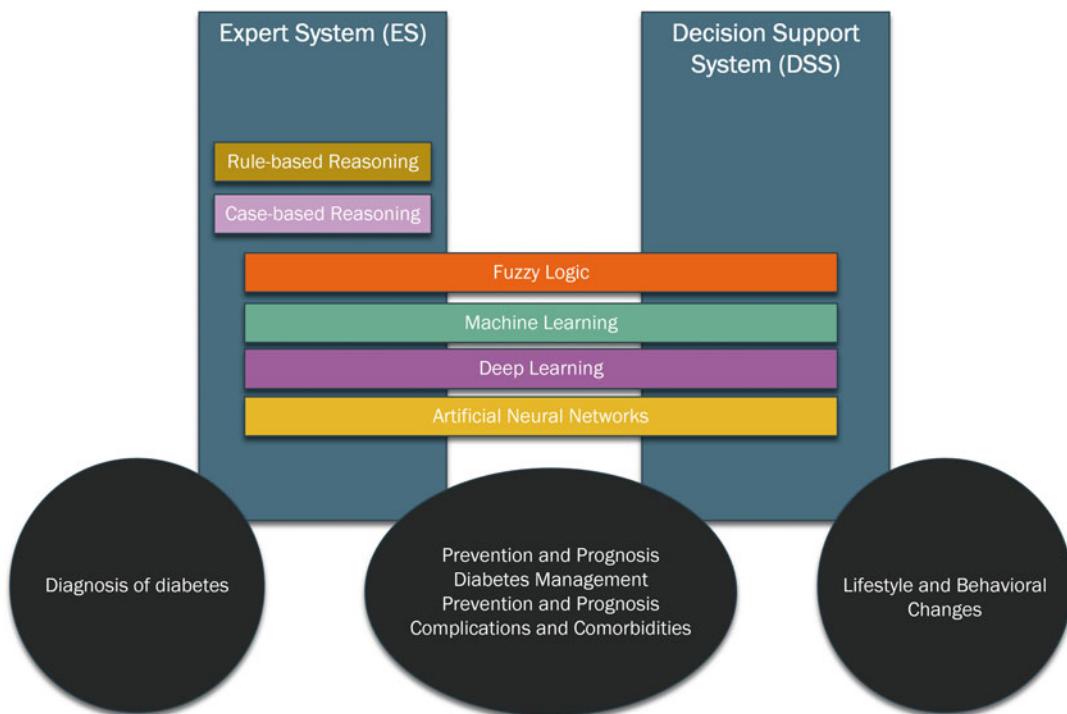


Fig. 1 An outline of AI impacting different areas of diabetes health care

patient's glycemic profile. Like ES, the DSS also utilize ML and DL models to predict, determine, and identify various trends in relevant data and base its decision on it.

Prevention and Prognosis

The prognosis of diabetes concerns the patient as well as the health care professionals. The knowledge of an individual developing habits or the physiological indicators that may lead to diabetes in future may tempt the patient to act proactively in order to save themselves from diabetes. Healthcare professionals may find this knowledge useful in recommending lifestyle changes and prescribing drugs to the individuals with diabetic tendencies.

T1D is developed in childhood or adolescent and its causes are often genetic. In almost all cases, it is impossible to predict the occurrence of T1D. On the other hand, T2D is more likely to happen in people after the age of 40. Though genetic factors could contribute greatly to the occurrence of T2D, majority of the time it is our lifestyle choices that triggers it. It is, therefore, possible to predict the risk of occurrence of T2D by analyzing different aspects of a person's lifestyle. This is a perfect ML problem that could be tackled by either classification or regression techniques. According to a review study, in 2018, the most common application of AI for T2D was screening and diagnosis in different stages [8]. In past ML models have been used to predict the risk of individuals developing T2D by using different demographic characteristics such as smoking, drug history, etc. [9]. AI-based models have also been used to predict the risk of diabetes from several other types of data like electronic health records (EHRs) and genetic data [10, 11]. Many other examples could be found in literature that predict the occurrence of T2D using AI and suggest different strategies for its prevention [12–14].

Diagnosis of Diabetes

As discussed earlier, almost half the diabetes cases around the world remain undiagnosed. What is even more alarming, a lot of the diagnosed cases

are wrong diagnosis. The danger of misdiagnosis is that the patients are put on inappropriate treatment regimes. Moreover, misdiagnosis increases the risk of diabetic ketoacidosis among T1D patients by about 18% [15]. The reasons for this could be many. Ranging from the scarcity of health care facilities to the accuracy of the diagnostic equipment diabetes patients remain at the risk of a late diagnosis, misdiagnosis, or no diagnosis at all. Early and accurate detection of diabetes can save many lives. AI-based tools can improve the situation in this area of diabetes health care. From labeled data of correctly diagnosed patients in the past, ML techniques may efficiently recognize whether a person suffers from diabetes or not. Moreover, diagnosis of comorbidities developed with diabetes through AI is also a trending area of research with diseases like diabetic retinopathy being diagnosed by various research studies [16, 17].

T1D is often diagnosed at an early age. T2D on the other hand may develop in the later part of life. T2D patients often develop prediabetes before they are diagnosed with diabetes. It means the rise in BG levels above normal range but not enough to be labeled as diabetes. Prediabetes may go unnoticed for years until serious issues arise. Studies have tried to diagnose prediabetes by using ML techniques trained on diabetes-related variables like age, family history, weight, and body mass index [18, 19]. Other studies have used data like plasma glucose concentration, diastolic blood pressure, and body mass for the diagnosis of T2D [20]. Screening of gestational diabetes has also been the topic of multiple studies [21, 22]. Whereas data like age, fasting BG, and mRNA have been used for training the ML models.

Diabetes Management

Diabetes management broadly means measures taken to keep glycemic profile of a diabetic patient within normal range. In the context of AI, diabetes management is a group of different tasks performed to keep the BG profile of patient inside healthy limits. AI-based diabetes management systems may use individual tasks or a

combination of such tasks to reach this goal. Blood glucose control strategies, BG prediction, detection of adverse glycemic events, insulin bolus calculators, and advisory systems and detection of meals, exercise, and faults [23] are all tasks that can be performed by an AI-based diabetes management system.

BG control strategies emerged as an area of study in diabetes management that is associated with the concept of artificial pancreas (AP). An AP's closed loop consists of a CGM, a control algorithm, and an insulin delivery system. Control engineering theory have been employed in giving shape to many of such control algorithms and up to a respectable accuracy. AI, however, is proving to be an improvement on these traditional control methods with techniques like ANN, fuzzy logic, and reinforcement learning. BG prediction is another task that AI is performing efficiently. Training ML models with BG values along with their associated time stamps can result in the forecast of future BG values. By forecasting the BG values of a diabetes patient at some time in future, a management system can help the patient in preparing for the possible occurrence of an adverse event. ML finds hidden nonlinear relationships in the training data and forms decisions based on these relationships. In case of diabetes, by routing these relationships we can detect the adverse glycemic events in present.

Insulin intake is a mandatory therapy for all T1Ds and some T2Ds. Insulin delivery to the body can be performed in multiple ways. Insulin taken at the time of meal intake is referred to as bolus insulin. This type of insulin tackles the sudden surge of BG caused by a meal. On the other hand, insulin taken in between the meals is called basal insulin. Figuring out the correct doses of bolus and basal insulin is a tricky process and something that a diabetic patient must always do correctly. AI has provided us with models that can compute the amount of insulin needed to keep the BG levels in normal range up to a certain level of accuracy [24, 25]. AI-based tools have been designed that detect events affecting the glycemic control of diabetic patients. These events may contain exercise, meals, or faults in the equipment [26–28]. The detection of such events may prove

vital in lifestyle decisions being suggested by an AI-based DSS.

Lifestyle Interventions/Behavioral Change

Decisions related to lifestyle and behaviors have a huge impact on the overall health of diabetics. These decisions may prevent diabetics from falling prey to any of the adverse glycemic events such as hypoglycemia and extreme hypoglycemia, etc. Moreover, if anticipated early, opting to a healthy lifestyle may prevent T2D from happening altogether. Diabetes DSS are platforms that help diabetics correct their lifestyle in order to avoid complications arising because of their illness and to live a normal life [29]. These lifestyle decisions may contain eating habits, exercise, sleeping habits, and insulin recommendations.

The DSS platforms often utilize ML models to compute measured impact of these decisions and give away suggestions to the diabetics so that they may avoid any unfavorable circumstance. With advances in smartphone technology, these DSS platforms are now more realizable than ever. A diabetic patient with an AI-driven DSS system on their smartphone is virtually under expert supervision at all times. A DSS running on a smartphone device in integration with CGM and PA sensors give birth to the idea of a potentially brilliant DSS. An apt DSS platform predicts the future course of BG values and then informs the diabetes patient about the action needed to avoid any trouble. These actions, based on the nature of adverse event, the type of diabetes, and variability of BG profile, may suggest different things to the patient such as the type of food consumption, the intensity of physical activity, and the dose of insulin.

Lifestyle and behavior changes suggested by the DSS along with reducing the risk of comorbidities and other adversities also ensure an improved life quality. The growth of diabetes is often linked with lifestyle behaviors that encourage obesity. Research studies have proposed strategies to prevent obesity endorsing habits that may cause diabetes later on

[30]. AI-based studies also encourage the patient's role in managing diabetes to meet the required targets. Indeed, a potent DSS can only be effective if the patient is willing to adopt the changes suggested by the system. Such self-management approaches are also referred to as patient-driven systems.

Risk Stratification

Risk stratification refers to quantifying the risk of an adverse event happening in present or is predicted to occur in future. It necessarily means that after detection and prediction, the AI algorithms are also capable of quantifying the risk a certain event poses to a diabetic patient. Risk stratification may prove to be a vital tool in scenarios where an extreme form of some adverse glycemic event poses a life threat to the diabetic patient. For instance, hypoglycemia is referred to as the decrease in BG below critical levels. Normally, when the BG levels drop down below 70 mg/dL a hypoglycemic event is diagnosed. An extreme case of hypoglycemia is reported when the BG levels drop below a level of 54 mg/dL. Extreme hypoglycemia can prove deadly if not dealt with in time. The concept of risk stratification in a scenario like this seems critical. If an AI-based system can predict the future occurrence of a hypoglycemic event and with it the risk of an extreme case, a complete hypoglycemic risk prediction system can be brought into realization.

Studies have measured the risk of hypoglycemia and certain other adverse glycemic events with the help of ML. Maniruzzaman et al. [31] has proposed a diabetes risk stratification study that uses the random forest model to assess the risk of diabetes. A risk stratification system that predicts the risk of gestational diabetes by using electronic health records was presented by Bilal et al. [32]. The application of risk stratification is not only limited to diabetes and its associated adverse events but also spans an entire group of comorbidities arising because of diabetes. Studies have used ML-based risk stratification strategies to quantify the risk of heart complications in diabetes patients [33–35].

Complications and Comorbidities

As discussed already, in many cases if the diabetes patients do not manage their illness properly, there is always a risk that other harmful conditions will start to appear in the body. These conditions that appear in the human body because of diabetes are called diabetic comorbidities [36, 37]. The most common of these comorbidities is hypertension, heart diseases, retinopathy, nephropathy, neuropathy, depression, autoimmune thyroiditis, foot damage, skin conditions, hearing impairments, and Alzheimer's disease. AI tools can be used for the detection and diagnosis of these comorbidities. Techniques like ANN and DL have made it possible to diagnose diseases from medical images and other types of medical signals. Many studies have used AI for the automatic screening of diabetic retinopathy [17, 38]. Other studies have utilized ML for the diagnosing and risk stratification of heart complications in diabetics [35]. Furthermore, studies have used AI in diagnosing cardiovascular complication that arise due to diabetes as well as the progression of diabetic kidney [39, 40]. Certain other works have diagnosed and predicted diabetic foot ulcers in diabetes patients [41, 42].

Discussion

AI has brought upon a revolution in health care that is set to transform almost all areas of the traditional health care system. In diabetes health care, AI is set to revolutionize major aspects of the system. From detection to prevention of diabetes, AI can be used to put a halt to the growing pandemic. This can result in saving billions of dollars spent each year in regard to direct medical cost and reduced productivity. For diabetic patients, AI can mean higher life expectancy, ease of access to professional-level advice, self-operating closed-loop insulin delivery systems, and peace of mind. AI-based blood control strategies are capable of replacing conventional control methods in AP. Moreover, these strategies have the potential to perform better than conventional methods. AI-based BG level prediction may pave way to

highly efficient DSS platforms that can provide professional-level advice based on future forecasting. An AI-based forecasting method can prove vital in the prevention of adverse glycemic events such as hypoglycemia. Such a method when combined with a risk stratification technique can prove to be a potent warning system in case of extreme adversity. AI-based DSS can be loaded with functionality to provide lifestyle advice to diabetic patients. Such DSS platforms can provide suggestions regarding diet, physical activity, and other habits of patients based on personalized AI models. AI-based screening of diabetic comorbidities can save time and money on both the patient and services side.

Conclusions

AI-based methods, with applications in almost every field of diabetes health care, have the potential to transform the scene of the entire system and have far-reaching results. It is now high time that the world recognizes AI in diabetes as a distinguished field of research. However, it is understood that AI is a field that will grow and reach perfection with time as the pool of available relevant data grows. Researchers need to explore new methods and continue experimenting with the existing AI models by coming up with hybrid techniques and cascaded models that may target any of the potential problems diabetics face.

References

- Jiang F, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol.* 2017;2(4): 230–43.
- Blair M. Diabetes mellitus review. *Urol Nurs.* 2016;36(1):27–36.
- Atkinson MA, Eisenbarth GS, Michels AW. Type 1 diabetes. *Lancet.* 2014;383(9911):69–82.
- Olokoba AB, Obateru OA, Olokoba LB. Type 2 diabetes mellitus: a review of current trends. *Oman Med J.* 2012;27(4):269.
- A. D. Association. Gestational diabetes mellitus. *Diabetes Care.* 2004;27(Suppl 1):s88–90.
- Bichindaritz I, Marling C. Case-based reasoning in the health sciences: what's next? *Artif Intell Med.* 2006;36(2):127–35.
- American Diabetes Association. Economic Costs of Diabetes in the U.S. in 2017. *Diabetes Care.* 2018;41(5):917–928. <https://doi.org/10.2337/dc18-0007>. Epub 2018 Mar 22. PMID: 29567642; PMCID: PMC5911784.
- Abhari S, Niakan Kalhori SR, Ebrahimi M, Hasannejadash H, Garavand A. Artificial intelligence applications in type 2 diabetes mellitus care: focus on machine learning methods. *Healthc Inform Res.* 2019;25(4):248–61.
- Ramezankhani A, Pournik O, Shahrabi J, Khalili D, Azizi F, Hadaegh F. Applying decision tree for identification of a low risk population for type 2 diabetes. *Tehran Lipid and Glucose Study. Diabetes Res Clin Pract.* 2014;105(3):391–8.
- Nguyen BP, et al. Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. *Comput Methods Prog Biomed.* 2019;182: 105055.
- Kim J, Kim J, Kwak MJ, Bajaj M. Genetic prediction of type 2 diabetes using deep neural network. *Clin Genet.* 2018;93(4):822–9.
- Perveen S, Shahbaz M, Keshavjee K, Guergachi A. Prognostic modeling and prevention of diabetes using machine learning technique. *Sci Rep.* 2019;9(1): 13805.
- Chaki J, Thillai Ganesh S, Cidham SK, Ananda Theertan S. Machine learning and artificial intelligence based diabetes mellitus detection and self-management: a systematic review. *J King Saud Univ – Comput Inf Sci.* 2020;in press.
- Birjais R, Mourya AK, Chauhan R, Kaur H. Prediction and diagnosis of future diabetes risk: a machine learning approach. *SN Appl Sci.* 2019;1(9):1112.
- Muñoz C, et al. Misdiagnosis and diabetic ketoacidosis at diagnosis of type 1 diabetes: patient and caregiver perspectives. *Clin Diabetes.* 2019;37(3):276LP–281.
- Padhy SK, Takkar B, Chawla R, Kumar A. Artificial intelligence in diabetic retinopathy: a natural step to the future. *Indian J Ophthalmol.* 2019;67(7):1004–9.
- Wolf RM, et al. Cost-effectiveness of autonomous point-of-care diabetic retinopathy screening for pediatric patients with diabetes. *JAMA Ophthalmol.* 2020;138:1063.
- Choi SB, et al. Screening for prediabetes using machine learning models. *Comput Math Methods Med.* 2014;2014:618976.
- Yu W, Liu T, Valdez R, Gwinn M, Khouri MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Med Inform Decis Mak.* 2010;10:16.
- Kumari VA, Chitra R. Classification of diabetes disease using support vector machine. *Int J Eng Res Appl.* 2013;3(2):1797–801.
- Yoffe L, et al. Early diagnosis of gestational diabetes mellitus using circulating microRNAs. *Eur J Endocrinol.* 2019;181(5):565–77.
- Shen J, et al. An innovative artificial intelligence-based app for the diagnosis of gestational diabetes mellitus

- (GDM-AI): development study. *J Med Internet Res.* 2020;22(9):e21573.
23. Contreras I, Vehi J. Artificial intelligence for diabetes management and decision support: literature review. *J Med Internet Res.* 2018;20(5):e10775.
24. Noaro G, Cappon G, Vettoretti M, Sparacino G, Favero SD, Facchinetto A. Machine-learning based model to improve insulin bolus calculation in type 1 diabetes therapy. *IEEE Trans Biomed Eng.* 2021;68(1):247–255. <https://doi.org/10.1109/TBME.2020.3004031>. Epub 2020 Dec 21. PMID: 32746033.
25. Shifrin M, Siegelmann H. Near-optimal insulin treatment for diabetes patients: a machine learning approach. *Artif Intell Med.* 2020;107:101917.
26. Vahedi MR, MacBride KB, Wunsik W, Kim Y, Fong C, Padilla AJ, Pourhomayoun M, Zhong A, Kulkarni S, Arunachalam S, Jiang B. Predicting glucose levels in patients with type1 diabetes based on physiological and activity data. In Proceedings of the 8th ACM MobiHoc 2018 Workshop on Pervasive Wireless Healthcare Workshop 2018;26:1–5.
27. Aiello EM, Toffanin C, Messori M, Cobelli C, Magni L. Postprandial glucose regulation via KNN meal classification in type 1 diabetes. *IEEE Control Syst Lett.* 2018;3(2):230–5.
28. Reddy R, Resalat N, Wilson LM, Castle JR, El Youssef J, Jacobs PG. Prediction of hypoglycemia during aerobic exercise in adults with type 1 diabetes. *J Diabetes Sci Technol.* 2019;13(5):919–27.
29. Pérez-Gandia C, et al. Decision support in diabetes care: the challenge of supporting patients in their daily living using a mobile glucose predictor. *J Diabetes Sci Technol.* 2018;12(2):243–50.
30. Ashrafzadeh S, Hamdy O. Patient-driven diabetes care of the future in the technology era. *Cell Metab.* 2019;29(3):564–75. <https://doi.org/10.1016/j.cmet.2018.09.005>.
31. Maniruzzaman M, et al. Accurate diabetes risk stratification using machine learning: role of missing value and outliers. *J Med Syst.* 2018;42(5):92.
32. Mateen BA, David AL, Denaxas S. Electronic health records to predict gestational diabetes risk. *Trends Pharmacol Sci.* 2020;41(5):301–4.
33. Milner J, Monteiro S, Monteiro P, He M, Simpson C, Zaslavskiy M, Balazard F, Li L, Rousset A, Schopf S, Dellamonica D. P6420 Can machine learning help us improve risk stratification of diabetic patients with acute coronary syndromes? The answer will blow your mind. *Eur Heart J.* 2019;40(Suppl_1):ehz746–1014.
34. Beatrice R, et al. Abstract 15892: machine learning techniques for risk stratification of non-ST-elevation acute coronary syndrome: the role of diabetes and age. *Circulation.* 2017;136(Suppl_1):A15892.
35. Segar MW, et al. Machine learning to predict the risk of incident heart failure hospitalization among patients with diabetes: the WATCH-DM risk score. *Diabetes Care.* 2019;42(12):2298–306.
36. Katsiki N, Tousoulis D. Diabetes mellitus and comorbidities: a bad romance. *Hellenic J Cardiol.* Netherlands. 2020;61(1):23–5.
37. Braunwald E. Diabetes, heart failure, and renal dysfunction: the vicious circles. *Prog Cardiovasc Dis.* 2019;62(4):298–302.
38. Bellemo V, et al. Artificial intelligence screening for diabetic retinopathy: the real-world emerging application. *Curr Diab Rep.* 2019;19(9):72.
39. Dey D, et al. Artificial intelligence in cardiovascular imaging: JACC state-of-the-art review. *J Am Coll Cardiol.* 2019;73(11):1317–35.
40. Makino M, et al. Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning. *Sci Rep.* 2019;9(1):11862.
41. Goyal M, Reeves ND, Davison AK, Rajbhandari S, Spragg J, Yap MH. Dfunet: convolutional neural networks for diabetic foot ulcer classification. *IEEE Trans Emerg Top Comput Intell.* 2018;4(5):728–39.
42. Goyal M, Reeves ND, Rajbhandari S, Yap MH. Robust methods for real-time diabetic foot ulcer detection and localization on mobile devices. *IEEE J Biomed Health Informatics.* 2019;23(4):1730–41.



AIM in Primary Healthcare

52

Niklas Lidströmer, Joseph Davids, Harpreet S. Sood, and Hutan Ashrafiyan

Contents

Introduction	713
Opportunities of AI in Primary Care Include:	714
Shift of Balance in Healthcare	714
Electronic Health Records and Data Ownership	715
Global Macrotrends [1–3]	716
Symptom Checkers and Dissemination of Specialities	716

N. Lidströmer (✉)

Department of Women's and Children's Health, Karolinska
Institutet, Stockholm, Sweden
e-mail: niklas.lidstromer@ki.se; niklas@lidstromer.com

J. Davids

Imperial College London NHS Trust, London, UK

Hamlyn Centre for Robotics and Artificial Intelligence,
Department of Surgery and Cancer, Imperial College
London, London, UK

National Hospital for Neurology and Neurosurgery Queen
Square, London, UK

e-mail: j.davids@imperial.ac.uk

H. S. Sood

Health Education England, London, UK
e-mail: harpreet.sood@nhs.net

H. Ashrafiyan

Department of Surgery and Cancer, Imperial College
London NHS Trust, London, UK

Institute of Global Health Innovation, Imperial College
London, London, UK

Hamlyn Centre for Robotics and Artificial Intelligence,
Department of Surgery and Cancer, Imperial College
London, London, UK

e-mail: h.ashrafiyan@imperial.ac.uk

Altered Roles	717
Precision Medicine and Frontiers for AI	732
Legal and Regulatory Aspects	732
Privacy Concerns	732
Patient Safety	732
Medical Imaging Diagnostics and Radiology	732
Medical Informatics and Clinical Decision Support	733
Patient's Perspective	733
Gender Aspects	734
Point-of-Care Dermatology and Ophthalmology	734
Public Health Aspects on Primary Healthcare	734
AI for General Practice Management	734
Endocrinology and Diabetes	735
Cardiovascular and Respiratory Management	735
Cardiac	735
Hypertension	735
Respiratory	735
Chronic Neurological and Neuropsychiatric Disease Monitoring	736
Obstetrics, Pregnancy, and Pediatrics	736
Oncology	736
Conclusion	737
Cross-References	737
References	738

Abstract

Primary healthcare is a highly interesting generalist field in medicine. Over the coming years, this field will continue to profoundly benefit from AI in medicine, which will result in positive changes in the everyday lives of patients.

Medical specialist knowledge will reach out to primary healthcare settings, profoundly altering the whole referral system and its indications. Specialist domains will be distributed widely and remotely, as scientific advances will reach primary care patients and doctors more frequently, rapidly, and accurately, thus tilting the dependency balance in the patient–doctor relationship.

Personalized and precision healthcare will reach out to every clinic and patient, and nowhere will it be as obvious as in the primary healthcare setting. AI in primary care will also

speed up disease theranostics, which will impact management decisions. Decision support will be abundant for the GP and the patient. Patient power will likely see an increase as patients become more active, well-informed, and independent in information discovery and learning about their own disease, a trend that has already occurred in most developed economies. This trend will likely continue in emerging economies through AI-powered mHealth platforms thanks to the rise in smart phone technologies.

Many areas within primary care are entering a revolution: *Pharmacogenomics* will profoundly change the way we prescribe medications. All types of pattern recognition in image-based specialties will essentially strengthen their presence in the primary care clinic: radiology ranging from flat X-rays to ultrasonography, dermatology, pathology, and parts of ophthalmology and

scopic inspections, where other image pattern recognitions can be further expanded. Moreover, interpersonal psychotherapy, follow-ups, and compliance will be armored with surveillance, coaching, and instructing components.

Verily, in primary healthcare, the whole medical AI symphony will reach its soaring tutti and eventually the energetic confluence of all the thematic lines – incorporating the *Allkunstwerk* of AI in medicine.

Keywords

Artificial intelligence · Primary healthcare · General practice · Clinical decision support · Electronic health records · Precision medicine · Personalized medicine · Democratizing medicine · Telemedicine · Deep medicine

Introduction

Accenture estimates that the health AI market in a country like the USA will grow at a 40% annual rate, reaching \$6.6 billion by 2021 (pre-SARS-COV-2). Reports by the WHO and World Bank suggest that the globalization of health system economies will leverage preventative/predictive AI to shift the management decisions from treatment to prevention [1, 2]. This can mitigate the growing cost of noncommunicable diseases (NCDs) linked with high morbidity and mortality, which are estimated to cause a cumulative loss to global output of \$47 trillion between 2011 and 2030 [1, 2]. Moreover, the WHO estimates staffing shortages of up to 12.9 million by 2035 [1, 3]. To this end, the primary care physician will be at the heart of a new wave of exciting patient-focused AI-based healthcare innovation and transition bringing about effective transformation of healthcare delivery for the improvement of outcomes.

As the gatekeeper to secondary healthcare access, general practice or family medicine involves early interaction with the patient identifying their initial presenting complaint and assessment prior to referral for specialist intervention or emergency services. AI in medicine is expected to cohere into a burning focus inside the domains of primary healthcare [4, 5]. Hence, we will

summarize this chapter into a condensed but relevant and frontline overview of all the possibilities and ongoing developments within the field of primary care. As many aspects as possible will be elaborated on in order to cover the effect of how AI will impact and be impacted by primary healthcare. This chapter will review all areas, applications, specialities, and tools, which we believe will impact primary healthcare.

Several changes including the shift towards digitization and the transition to electronic health records for most societies will continue to evolve the doctor-patient interaction enabling patients to gain legally binding ownership of their data. In doing so, patients will also need to become more aware of the ethico-legal responsibilities that surround ownership of their personal data management and what constitutes mishandling of data and security breaches. This issue will require a security backed AI system that will support and alert the patient of such a breach or ideally prevent such breaches. This will be tied in with patient education about their condition, the right source of data, data integrity and data security, etc. It would not be surprising to see the emergence of a private patient owned health cloud in some countries, with perhaps even a global potential for regulated information discovery.

This central storage place with the patient at the center and with the invited professional human in the loop, likely in the form of the patient's own general practitioner, would enable a phenomenal overview and new insights. This setup will give patients and clinicians the opportunity to deliver personalized treatment, i.e., real precision medicine. No future primary healthcare would be complete without the in-depth integration of genomic data, which would bestow upon us the opportunity for real prevention. Equally, pharmacogenomics will give the patient the optimal drug immediately, instead of trial and error. The integration of microbiomics will also take the primary care dietician's role to an evolved level [6]. Deep diet will allow machine learning support for optimizing nutritional value, with significantly greater evidence-based backing. Hence, AI supported evidence-based medicine (EBM) in primary healthcare will facilitate effective multidisciplinary primary care management decisions.

A macrotrend with AI will be the moving of more specialist tasks and knowledge closer to the hands of the patient and their GP. Though, it can also be argued that in some situations the patient can directly seek the highest specialist competence, seemingly bypassing the GP to enter secondary care via emergency medicine, AI triaging systems supporting out of hours services will grow in prominence over the coming decades. As mentioned before, the role of the GP may evolve into taking on a more God-view holistic approach and subspecialist forms may even become technical with GPs programming management algorithms for AI systems to meta-learn. However, in a forest of narrow AI applications, it will still be continuously valuable to have an overviewing general practitioner who has the empathy and knowledge base to support the patient. Although in some economies, the time pressures and resource constraints limit such effective doctor-patient interaction. That said, there will always be room for deep empathy and counselling, and even thinking outside the black box in more intricate cases. In fact, more, not less, may be required from these primary care specialists.

In the discussion of AI in primary healthcare, it is a mistake to only refer to the changing landscape of the healthcare and situation in Europe or the USA – the effect will be as, if not more, dramatic in countries with emerging economies [1, 3]. Globally, we will see the dawn of primary healthcare with precision, individualism, specialist assets, prevention, and broader patient engagement. The enormous spread of mobile smart devices will further fuel this inevitable major global trend.

Given the fact that most medical conditions, including those requiring highly advanced specialist treatment, e.g., in hematology, immunology, interventional radiology, or oncology, all enter the hospital labyrinth through the gate of primary healthcare, the GP will remain in an advantageous seat to leverage big data to generate insights that will support patient management. They usually oversee and know a lot about the patient's family, environment, work, and other circumstances – a knowledge base greater than what could easily be achieved by a naïve AI. GPs will prove crucial in

improving explainability and overcoming the "black box" nature of an AI, thus helping to improve the limited capability of some narrow algorithms.

In summary, the primary healthcare specialist sits at the crossroads of human destinies, relations, subtle facts, and almost literary conditions, which cannot always be fully transcribed, no matter how delicately a medical record text is composed. It is a matter of deep empathy, knowledge, and dedication in the care of a fellow human being. AI will hopefully help to augment and progress the doctor-patient relationship, while allowing patients to become independent in understanding their disease better and co-managing it with their GPs.

Opportunities of AI in Primary Care Include:

- Gathering and synthesizing disparate patient data leading to the efficient integration of health information (text, images, and knowledge) – also increasingly relevant with data from wearables, sensors, and remote monitoring
- Increased efficiency of establishing diagnoses and treatment
- Predicting and possibly averting complications and reducing unnecessary services/costs
- Saving time on administrative tasks
- Allowing each member of the care team to perform at the highest level

Shift of Balance in Healthcare

In most developed and developing economies, multiple drivers such as patient demographics, gender differences, availability of services and financial support limit the balance of patient-primary care practitioner healthcare provision [7]. The advent of technology and the internet is helping to shift the balance more towards patient self-empowered healthcare delivery. Patients are researching their conditions online and will usually come into a general practice consultation very much more well-informed.

The difficulty is ensuring that the information for the patient is the right information provided by reputable sources. Over the coming years, artificial intelligence can help in the source verification process to facilitate this process for primary care patients. The shift in balance of healthcare information delivery will no longer be focused on the GP giving the information to the patient, but the patient taking charge of researching their own condition. The GP's role as well as offering treatment will shift towards empowering the patient to independently interpret the growing complexities of available data, and this can be achieved using AI platforms.

Electronic Health Records and Data Ownership

Sinsky and colleagues conducted a study in 2016 which found that physicians spent 27% of their office day on direct clinical face time with their patients and 49.2% on electronic hospital records and desk work [8]. When in the examination room with patients, physicians spent 52.9% of their time on EHR and other work. Physicians who used documentation support such as dictation assistance or medical scribe services engaged in more direct face time with patients than those who did not use these services. In addition, increased AI usage in medicine not only reduces manual labor and frees up the primary care physician's time but also increases productivity, precision, and efficacy.

In many developed and some emerging health economies, the shift towards digitization and the transition into the electronic health record has to coincide with schemes that improve the patient's understanding about the legally binding responsibilities associated with them owning and controlling their own data. Moreover, there needs to be an obligation to support the understanding of the ethico-legal responsibilities that surround the patient's owned and personalized data management [9–11]. This constitutes the education of data mishandling and security breaches, which is difficult for the digitally naïve to appreciate. This issue will require a cybersecurity-aware AI system primed on healthcare data breaches, and a policy that will support and alert the patient of

such a breach. This will be tied in with patient education about their condition, the right source of data, data integrity, and information governance, as well as data security, etc. [9–11]. Patients will need to know about ransomware and adversarial attacks or be supported in multiple methods to help them become aware of this or detect this if they are to own and manage their data. AI platforms that take into account encryption systems could support these issues.

Proponents have leaned towards patient data being co-located within a central private cloud infrastructure and co-controlled and accessible by both the patient and the primary healthcare provider [12, 13]. While others desire a more decentralized system where current technologies mean that various forms of data and the storage of all the big-omics data, e.g., genomics and gut microbiome, and all health records including radiological, dermatological, pathological, and other clinical images, laboratory results and the written record have to be handled effectively. Relatedly, such data has to be made rapidly available to the physician in their consultation or provided as insights and dashboards for rapid decision support [14]. This would mean that patients cannot attend a clinic forgetting to take crucial data with them, which usually leads to delays in consultation or a feeling that the visit to the clinic was fruitless for the patient. It is for this reason that one proposed efficient delivery, where the storage system is a digital server or personalized secured cloud-based infrastructure.

Those deeply in-favor of the patient-centric approaches to data management assert that clinical data is not complete without including possible verbatim comments made by the patient for medicolegal purposes. How this patient information is captured in the initial consultation may need to shift towards a more digital record using AI processing of audio-centric recordings. AI systems that can auto-transcribe interactions between the GP and their patient is something that has also been reported [15]. Platforms that keep the GP and patient in the loop about any updates including results from recent investigations will become more prominent in a co-owned system.

The lack of communication between electronic health records is also a major issue when it comes

to information transfer across primary care systems, or between primary and secondary care either cross-border or locally. Although internet platforms currently allow email encryption for data transfer. Usually patients end up receiving printed paper copies of their data in a non-environmentally friendly and unsecure manner. AI systems that can encode information from one data system to another data system for rapid retrieval and that take into account intersystem security protocols will become invaluable. Decentralized systems aim to deliver potential safeguards, but it remains to be seen whether this will be the case and ubiquitously utilized in primary care.

Though it should be added that security codecs are not necessarily AI, but when a specific recognition is included, e.g., by a translation or text expansion/interpretation step, AI would be justified.

Global Macrotrends [1–3]

Tourism and migration have been linked to a trend towards cross-border globalization of healthcare being powered by the internet and tele-health platforms, which are simplifying the move towards the cross-border delivery of care. These trends are going to be augmented by AI for the benefit of effective primary healthcare delivery. It is not infeasible that migrant patients will be able to receive integrated healthcare with information management between their primary care physicians in their home countries and specialists in their country of residence. The financing of this type of system adds additional complexities that will require cross-border treaties and policies to be established to facilitate this, taking into account the language barriers and the need to have AI systems that can aid translation. It did, however, become clear that during the 2020 SARS-CoV-2 pandemic, cross-border dissemination of information was deemed feasible to enable vaccine development, all of which was powered by machine learning techniques. However, this also required a coordinated effort within local community healthcare teams to facilitate vaccinations using AI platforms to support the effort. In summary:

The impact of AI on primary care is linked to patient treatment decision support systems, public health surveillance systems, and clinical trial management and follow-up.

Symptom Checkers and Dissemination of Specialities

Several AI-based symptom checkers have been developed with the aim of improving the accuracy and the reliability of any advice, guidance, or signposting. In a vignette-based audit of 23 symptom checkers available for use by the public, the appropriate triage advice was considered to have been provided in 57% of evaluations, rising to 80% for emergency cases [16]. These numbers must, however, be put in perspective – how good is a human expert on average when confronted with the same signs, opinions, or measurements?

Unfortunately, a major problem with most symptom checkers is that they are risk averse, encouraging users to seek help from their GP for conditions when self-care would be perfectly reasonable and safe. Striking a sensible balance between missing a serious condition and overwhelming GP workload is tricky. As many symptoms presenting to GPs are self-limiting, one option to consider is the development of AI tools that enable certain symptoms to be monitored over time [17]. AI systems can also be fine-tuned to improve diagnostic yield and optimize their screening potential to identify when to involve the general practitioner.

In some areas where primary care services offer remote care to the patient detached from urban regions and hospitals, the consultant/specialist might be required to intervene remotely to action emergencies occurring in primary care. Newer referral principles that can facilitate this will be augmented by artificial intelligence. Various specialist services including dermatology and pathology, etc. may greatly benefit from augmented diagnostic approaches using trained machine learning models to direct diagnostics ensuring prompt referral for specialist support is achieved [18, 19]. Smartphones can now be used as the diagnostic tool of choice with peripherals that can image the ear drum, or in the case of dermatology, a skin lesion [18–20]. These machine learning

algorithms will tirelessly classify these lesions and provide feedback to the patient and facilitate remote specialist input from the clinician. Doing so frees up hospital clinic slots, bed space, and emergency rooms for the admission of critically unwell patients. There are also developments where a smartphone can be exploited to make high-resolution retinal images, e.g., for screening for and longitudinal monitoring of diabetic retinopathy (see *smartphone retinal camera*, in Table 2).

Altered Roles

The evolution of the roles of nurses, physiotherapists, psychologists, curators, and the effect on the whole primary care team mean that these specialists will benefit from artificial intelligence systems to help manage patients with evolving expectations within a digitized economy. On one hand, there is also redistribution of roles in primary care, with concepts such as digital first

using e-consult/telephone triaging replacing in-person patient appointments. Others mention that huge swathes of the population are yet to embrace the use of smartphones in some parts of the world. However, a significant proportion of the global population have smartphones making AI-based mHealth systems an ideal approach to support the masses.

The transition of a patient from secondary care back to the primary care environment usually requires input from a multidisciplinary team of, for example, community physiotherapists and other clinicians. AI platforms can, and in some instances already do, coordinate this process within the community. Figure 1 summarizes a great many medical areas that are benefitting from AI for primary care. Table 1 summarizes the machine learning models in this space, and Table 2 presents a handful of the companies working on AI for primary care. We next discuss AI for other specialty body systems that usually require primary care input.

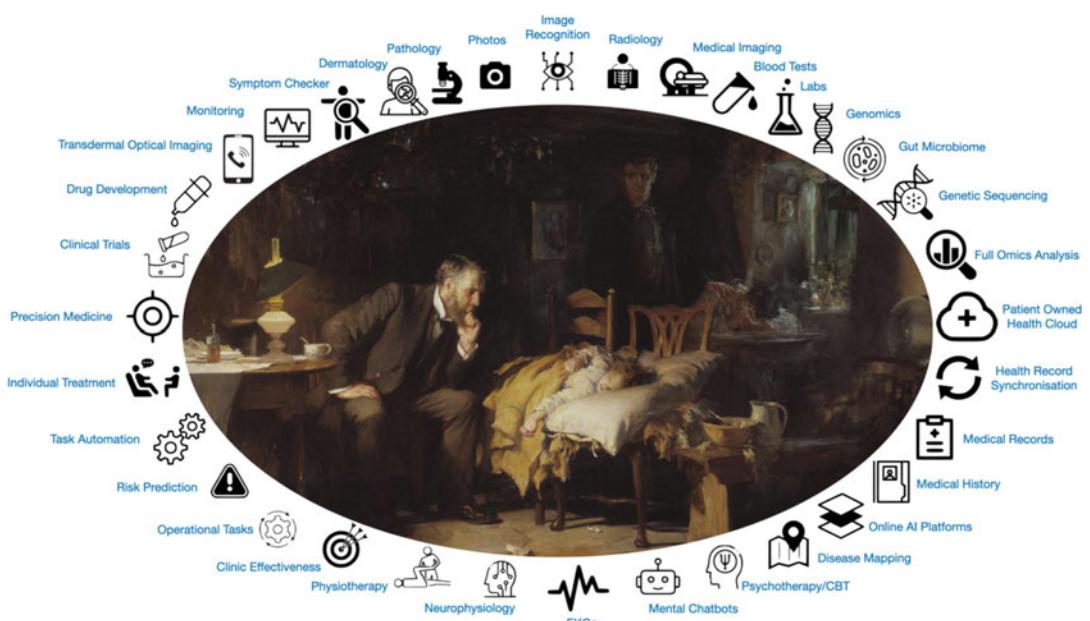


Fig. 1 The archipelago of artificial intelligence applications in primary healthcare at the time of the fourth industrial revolution of the 2020s. In the center *The Doctor*, exhibited 1861. Sir Luke Fildes 1843–1927. Tate/Tate Image. Central Image permission granted by Original Owners. Adaptation and re-imagination by Dr. Niklas Lidströmer. It demonstrates a striking connection between what doctors could do, then the evolution up to where we

are and what patients themselves and their doctors will soon be able to accomplish bound by developments in machine learning. The icons present a summary of the multidisciplinary aspects of general practice and subspecialties that interface with primary care where developments from AI are being made to support optimal patient care

Table 1 List of global companies working on AI platforms that will benefit primary care

Company	Main area	Actions	Reference site	Location
PathAI	Pathology	Reduced error. Personalized care	pathai.com	Cambridge, Massachusetts
Buoy Health	Symptom checker	Chatbot. Diagnosis. Treatment guidance	buoyhealth.com	Boston, Massachusetts
Enlitic	Radiology images, blood tests, EKGs, genomics, patient medical history	Multifaceted analysis platform	enlitic.com	San Francisco, California
Freenome	Screenings, diagnostic tests, and blood work	Cancer testing. Detection in the earliest stages	freenome.com	San Francisco, California
Beth Israel Deaconess Medical Center	Microscope bacteria scanning	Very early infection diagnostics	bidmc.org	Boston, Massachusetts
Zebra Medical Vision	Radiology assistant	Automatic analysis for clinical findings	zebra-med.com	Shefayim, Israel
Bioxcel Therapeutics	Drug development in immuno-oncology and neuroscience	AI to find new applications for existing drugs and identify new patients	bioceltherapeutics.com	New Haven, Connecticut
Berg Health	Disease mapping of common and rare diseases	Discovery and development of breakthrough medicines. Found new links between chemicals and Parkinson with AI	berghealth.com	Framingham, Massachusetts
Xtalpi	AI + cloud + quantum physics. The ID4 platform	Predicts the chemical and pharmaceutical properties of small-molecule candidates for drug design and development	xtalpi.com	Cambridge, Massachusetts
Atomwise	Focus on serious diseases, including Ebola and multiple sclerosis. AtomNet	Screens with AI 10–20 million genetic compounds per day. Results 100 times faster than traditional pharmaceutical companies	atomwise.com	San Francisco, California

Deep genomics	Drug candidate identification	Neuromuscular and neurodegenerative disorders	deepgenomics.com	Genomics will be a naturally integral part of all future primary healthcare	Toronto, Canada
Benevolentai	Right treatment to the right patients at the right time	Better target selection and provide previously undiscovered insights	benevolent.com	Drug licensing and easily transportable medicines for rare diseases may be integrated into primary healthcare	London, England
Olive	Automation of healthcare's most repetitive tasks, freeing up administrators to work on higher-level ones	Eligibility checks to unadjudicated claims and data migrations so staffers can focus on providing better patient service	oliveai.com	Automation tools will integrate into the plethora of softwares and tools in primary healthcare	Columbus, Ohio
Qventus	Platform solving operational tasks, e.g., related to emergency rooms and patient safety	Prioritizing patient illness/injury, tracking of clinic waiting times, charting the fastest ambulance routes	qventus.com	A multitude of these tools have relevance to primary healthcare	Mountain View, California
Babylon hHealth	Personalized and interactive healthcare	Anytime face-to-face appointments with doctors and/or chatbots	babylonhealth.com	Solutions such as this one will integrate with physical primary healthcare clinics	New York, New York
Cloudmedx	Generating insights for improving patient journeys throughout the healthcare system	Clinics management of patient data and clinical history with predictive analytics to intervene at critical junctures in the patient care experience	cloudmedxhealth.com	Primary healthcare providers could use these insights to likely more efficiently move patients through the system with less of the traditional confusion	San Francisco, California
Cleveland Clinic	With IBM. Gathering information on trillions of administrative and health record data points to streamline the patient experience	AI and vast databases help personalize healthcare plans on an individual basis	clevelandclinic.org	As in the Cleveland clinic, primary healthcare clinics throughout the Worlds will benefit from insight from the marriage of AI and gigantic medical records and all other medical databases	Cleveland, Ohio
Johns Hopkins Hospital	With GE. Predictive AI techniques to improve the efficiency of patient operational flow	Hospital effectiveness program resulted in, e.g., augmented patient admission and discharge before noon	hopkinsmedicine.org	Applicable also to clinic effectiveness in primary healthcare	Baltimore, Maryland
Tempus	World's largest collection of clinical and molecular data in order to personalize healthcare treatments	Collection and analyses of data in everything from genetic sequencing to image recognition, that can give physicians better insights into treatments and cures	tempus.com	A massive data library for personalized care would also benefit primary healthcare	Chicago, Illinois

(continued)

Table 1 (continued)

Company	Main area	Actions	Reference site	For primary healthcare	Location
Kensci	Prediction of clinical, financial, and operational risk by taking data from existing sources	Foretelling anything from who might get sick to what's driving up a clinic's healthcare costs	kensci.com	Prediction tools of various sorts relevant for primary healthcare to best use the means and get the best healthcare for the money	Seattle, Washington
Proscia	Detection of patterns in cancer cells. Connection of data points that support cancer discovery and treatment	Helps pathology labs eliminate bottlenecks in data management	proscia.com	Digitalized pathology will move closer to primary healthcare in ways, also involving cancer pathology and oncology and its complex plethora of best treatment considerations	Philadelphia, Pennsylvania
H2O.AI	Data analysis of a healthcare system to mine, automate and predict processes	Prediction of ICU transfers, improve clinical workflows, or pinpoint a patient's risk of hospital-acquired infections	h2o.ai	Prediction and detection of sepsis, which ultimately reduces death rates, are clearly advantageous in primary healthcare	Mountain View, California
IBM	Harnessing data to optimize hospital efficiency, better engagement with patients and improve treatment	Watson is currently applying its skills to everything from personalized health plans to genetic testing results and catching early signs of disease	ibm.com/watson-health	Integration of Watson and other AI tools into primary healthcare will be a tangible game changer	Armonk, New York
Google DeepMind Health	In use by hospitals all over the world to help move patients from testing to treatment more efficiently	Doctor alerts when a patient's health deteriorates. Can help in the diagnosis of ailments by combining its massive dataset for comparable symptoms	deepmind.com	Also by collecting symptoms of a patient in primary healthcare and putting them into the DeepMind platform, doctors will be able to diagnose more quickly and effectively, especially in more rare and complicated cases	London, England
Icarbonx	Uses AI and big data to look more closely at human life characteristics in a way described as "digital life"	Analysis of health and actions of human beings in a "carbon cloud," with intentions of managing "all aspects of health"	icarbonx.com	Gathering of data to better classify symptoms, develop treatment options, in various solutions and forms, will most likely be integrated to the new primary healthcare	Shenzhen, China
Nuralogix	Transdermal Optical Imaging	Most vital signs given with the smartphone camera	nuralogix.ai	Of great use for screening not only in clinics but everywhere and in the hands of the patients	Toronto, Canada

BeeHealthy by Mehiläinen	Medical platform for online clinics. Offering over branding	Digital clinic with triage, online bookings, and digital patient journey	beehealthy.com	The AI-embedded online primary healthcare will be open 24/7, also offering access remotely, and where healthcare will be subject to “patient experience” monitoring	Helsinki, Finland
DayTwo	Gut microbiome. Evidence-based, precision nutrition	Deep diet – EBM personalized diet in diabetes, obesity, metabolic and gastrointestinal disorders	daytwo.com	Diabetes, metabolic and GI disorders are very large groups – the microbiome will be an integral part of evidence-based primary healthcare	Tel Aviv-Yafo, Israel
MyBioma	Gut microbiome	As above	mybioma.com	As above	Vienna, Austria
Kaiser Permanente Foundation	Electronic health record synchronization	An example of an organization which has integrated the EHRs and opened them for its patients	healthy.kaiserpermanente.org	An enormous global and local problem is the lack of integration and compatibility between EHRs	Oakland, California
Aysa	Dermatology	Analysis of skin pictures	askaysa.com	Of great relevance. Patients and GPs will be guided to the fastest route	Buenos Aires, Argentina
Mindler	Psychotherapy	Online CBT	mindler.se	Great relevance	Stockholm, Sweden
Betterhelp	Psychotherapy	Online CBT	betterhelp.com	Great relevance	Mountain View, California
SwordHealth	Physiotherapy	AI-embedded physiotherapy	swordhealth.com	Great relevance	Porto, Portugal
ADNtro	Complete genetic analysis	Genomics. At the core of preventative and individualized precision medicine	adntro.com	Genomics will be an integral part of future primary healthcare	Mallorca, Spain
MissionBio	Multi-omics screening	Interplay of genotype and phenotype	missionbio.com	Future primary healthcare with multi-omics with simultaneous detection of SNVs, CNVs, and protein at the single-cell level	San Francisco, California

Table 2 List of machine learning algorithms applied to specialties that interface or directly affect primary care patient management

	Area	Action	Machine learning algorithm	Explanation	Example	Reference
	Image recognition	A ground principle applicable to a long range of medical images; x-rays, photos of skin, eardrums, throats, faces, pathology slides, etc. Timing with the dramatically increased use of electronic medical records and diagnostic imaging	Convolutional neural network is one of the most prominent approaches	Medical image recognition or classification, e.g., based on deep features extracted by deep models and statistic feature fusion with multilayer perceptions	Filipovych et al. (2011) provide a recent example of semi-supervised learning to image recognition. There is a tremendous amount of successful ML in image recognition examples	[51, 52]
	Radiology images	AI will move radiology much closer to primary healthcare, enabling patient and GP insight into a long range of radiological images subtypes	Both unsupervised and semi-supervised learning are frequently employed in cluster analysis. The Cruz-Roa image recognition algorithm, employs supervised learning, in which the algorithm is presented with a training set of instances that provide, for each instance, both the covariates and the ground truth	ML algorithms have proved equal or superior to humans in interpreting X-ray and MRI images and slides. For example, Cruz-Roa et al. (2017) demonstrates that a trained ML algorithm achieves near-perfect detection of breast cancer at a microscopic level	Segar et al. (2020) provide a recent application of unsupervised learning to cluster analysis of heart failure. Another example from lung CT images: automated detection of pulmonary nodules: with morphologic matching algorithms	[53–55]
	X-rays, CT, MR, ultrasounds	All types of radiology, auto-diagnostics, radiology assistants	Supervised learning algorithms	Decreasing the burden of routine tasks, spreading radiological competence to primary healthcare	For example, Cruz-Roa et al. (2017) demonstrates that a trained ML algorithm achieves near-perfect detection of breast cancer at a microscopic level. Their algorithm, like other image recognition algorithms, employs supervised learning, in which the algorithm is presented with a training set of instances that provide, for each instance, both the covariates and the ground truth	[53]

	Blood tests	Quick and accurate medical diagnoses	Support vector machines (SVM), with the scikit-learn implementation SVC, which is based on the libsvm library	With respect to the tunable parameters, we experimented with both linear and radial basis kernels. The Γ parameter was calculated by the heuristic 1/number of attributes. The penalty parameter C was tuned using internal cross-validation in the training set	E.g., in hematological diagnoses	[56]
	Labs	To safely reduce the laboratory test ordering	Algorithm-based decision rule methodology	The number of tests can be reduced while missing critical values in only a small fraction of patients. Testing algorithms such as these can be used to reduce laboratory test ordering without compromising the quality of patient care	Decision rules to define appropriate intervals at which repeat tests might be indicated for commonly ordered laboratory tests for hospitalized patients	[57]
	Genomics	Analysis of full of genome sequencing	Clustering algorithms	Microarray technology has enabled measuring expression of thousands of genes simultaneously, hence the use of clustering algorithms	To discriminate pathologies based on their differential patterns of gene expression	[58]
	Gut microbiome	“Deep diet,” evidence-based diet for specific conditions, e.g., diabetes, obesity, metabolic disorders, etc.	E.g., the RDP classifier algorithm and gut microbiota-based random forest algorithms	Often a combination of several sets and types of algorithms; e.g., the random forest algorithm can be used to create a classification model	EBM approach to weight loss, but gut microbiota-based algorithms can also be used in the prediction of metachronous adenoma in colorectal cancer patients following surgery	[59]
	Genetic sequencing	It can become commonplace in primary healthcare with genetic sequencing, integrated in the medical records	GECKO is a genetic algorithm to classify and explore high throughput sequencing data	Genetic algorithms and feature selection to comprehensively explore massive volumes of sequencing data to classify and discover new sequences of interest	GECKO for GEnetic Classification using k-mer Optimization is effective at classifying and extracting meaningful sequences from multiple types of sequencing approaches including mRNA, microRNA, and DNA methlyome data	[60]

(continued)

Table 2 (continued)

	Area	Action	Machine learning algorithm	Explanation	Example	Reference
	Full omics analysis	In the future of primary healthcare not only genomics will be present but also other major factors, commonly called “omics”, e.g., genomics, transcriptomics, methylomics, proteomics, microbiomics	Multi-omics and multi-view clustering algorithms	Recent high throughput experimental methods have been used to collect large biomedical omics datasets. Different cancer types can be used as benchmarks	Clustering of single omic datasets has proven invaluable for biological and medical research. The decreasing cost and development of additional high throughput methods now enable measurement of multi-omic data	[6]
	Patient owned health cloud	As growing volumes of patient data are stored on externally hosted platforms, worries are mounting among patient advocates that patient data might be at risk of privacy breaches. It can be argued that the natural owner of the entire data related to a patient is the <i>patient</i>	Eric Topol (director and founder of Scripps Research Translational Institute in San Diego, California) noted that advances in mathematics, computing power, cloud computing, and algorithm design have accelerated the development of methods that can be used to analyze, interpret, and make predictions using these data sources	AI has the potential to transform the delivery of healthcare in a large national system, from streamlining workflow processes to improving the accuracy of diagnosis and personalizing treatment, as well as helping staff to work more efficiently and effectively	Louisiana's largest health system, Ochsner Health System, used advanced machine learning algorithms to create a predictive model leveraging Epic's machine learning platform powered by Microsoft Azure to accurately predict patient deterioration hours before an adverse event	[61]
	Health record synchronization	Currently electronic health records are spread out and cannot interact with each other – this is one the most serious global healthcare conundrums. With synchrony both the patient and primary healthcare will see massive benefits	A synchronized medical record would be more complete and more useful for deep learning CNNs. Moreover such an omni-EHR would contain omics data, imagery, smartphone and wearable vital sign monitoring data, etc.	Patient data synchronization processes in a continuity of care environment is pivotal, locally and globally – to ensure exchange and access to EHR data	An XML-based synchronization model that is portable and independent of specific medical data models. The implemented platform consists of several servers, of local network clients, of workstations running user's interfaces, and of data exchange and synchronization tools	[62]

	Medical records	Medical records are seldom ergonomic, and there have even been publications about doctor and nurse burnouts, “caused by the EHRs.” The EHR of the future primary healthcare must be improved, AI-embedded, and easily managed	Skim-gram algorithm	To improve the management ability of patient information and establish a feasible electronic medical record (EMR) management system, combined with the characteristics of e-commerce	The design of electronic medical records system using Skip-gram algorithm. Moreover cross-system compatibility is pivotal – Prognos (see below) uses machine learning to run its software which analyses electronic medical records from various hospitals and healthcare systems	[63]
	Medical history	A patient’s EHRs, if assembled, and synchronized and containing a wide range of data, is a veritable goldmine for medical AI	Pieces™ by Pieces Tech, KenSci, ZEUS by CareScore	Predictive Analytics (e.g., Pieces™ or KenSci), Diagnostic Analytics (e.g., Prognos), Prescriptive Analytics (e.g., ZEUS)	Based on a patient’s medical history, also Random Forest can be used to predicting the risk of disease	[63]
	Online AI platforms	Increased demand from primary healthcare, online telemedicine has led to several AI-embedded medical platforms on the market	Pieces™ Decision Sciences (DS) by Pieces Tech or KenSci or Prognos (the latter trained on a database of clinical diagnostics information with data for 50 diseases and its 1,000 algorithms are trained to analyze over 14 billion medical records for 180 million patients)	Machine learning and natural language processing can build a software platform to interpret patient data and recommend personalized treatment approaches	Pieces™ and KenSci are examples of predictive analytics. KenSci’s platform integrates with clinical systems such as electronic medical records (EMRs) so recommendations may be implemented by healthcare teams. See also Google’s site (referred on the side) on build-in cloud AI-platform algorithms	[64]
	Disease mapping	Spatiotemporal disease mapping models are a popular tool to describe the pattern of disease counts	Counts are usually formulated in a hierarchical Bayesian framework with latent Gaussian model. Or with INLA	Computationally expensive Markov chain Monte Carlo algorithm can be used for parameter estimation which might induce a large Monte Carlo error	An alternative method using integrated nested Laplace approximations (INLA) has recently been proposed	[65]

(continued)

(continued)

Area	Action	Machine learning algorithm	Explanation	Example	Reference
 Psychotherapy/ CBT	In primary healthcare, there is a shortage of educated and available human psychotherapist, which can offer, e.g., cognitive behavioral therapy (CBT). There is likely an over-prescription of selective serotonin reuptake inhibitors (SSRI), since the available therapists are often fewer than the demand. Patients are left without proper medication follow-up or CBT, despite the well-published benefits of therapy	There are two broad types of supervised and unsupervised learning (Friedman et al., 2001), relevant here. Supervised learning aims to solve prediction problems. For example, patients' features can be used to forecast the future development of clinically relevant events or measures. Unsupervised learning aims to solve classification problems, such as the discovery of emergent clinical phenotypes or processes	AI-empowered and/or online psychotherapy; see also chatbots below, will likely help the situation with a shortage of therapist or can help triage or classify of definite problem, so that therapist can focus more on the defined core of a problem. Not to forget to mention the valuable focus of human interaction per se; listening and what Eric Topol coined <i>Deep Empathy</i>	Supervised learning examples of clinically relevant events (e.g., treatment dropout) and of measures (e.g., symptom severity). Unsupervised learning examples of discovery of emergent clinical phenotypes (e.g., patients with similar features) or processes (e.g., interpersonal patterns encoded in textual, acoustic, image or biometric data)	[66]
 Mental chatbots	To interact, guide, and/or triage patients in psychiatry	High-level natural language understanding (NLU), and emotion recognition based on multimodal approach	Studies have even shown, that in some cases, especially of very sensitive problem nature, a patient can feel more comfortable speaking to a chatbot, at least during a certain period of the therapy	A chatbot for psychiatric counselling in mental healthcare service based on emotional dialogue analysis and sentence generation	[67]
 EKGs	Automatic 12-lead EKG interpretation overriding the present EKG interpretations	Deep neural network (DNN)	Models composed of stacked transformations that learn tasks by examples	Telehealth Network of Minas Gerais, collected under the scope of the CODE (Clinical Outcomes in Digital Electrocadiology) study.	[39]
 Neurophysiology	The specialty of neurophysiology will move closer to primary healthcare with the recent AI-advances, and which will change referral patterns	Three major categories for seizure detectors; EEG-based seizure-event detectors, EEG-based seizure-onset detectors, and EEG/ECG-based seizure-onset detectors	Epilepsy is a chronic disorder of the CNS that predisposes individuals to recurrent seizures and is used as an example here. Neurophysiology contains a plethora of disorder specific algorithms	Seizure detection algorithms based on analysis of EEG and ECG signals. Computerized seizure detection algorithms can enable alerting systems that may decrease the harm of the seizures	[68]

	Physiotherapy	A veritable explosion of smart devices and especially wearables, and mHealth in general, will fuse casual wearing, sporting, and medical monitoring with primary healthcare, i.e., the semantic definition of healthcare will be widened	Data can be collected in a clinic under supervision and can be used to train and validate ML algorithms. Validated algorithms will then be used to assess home physiotherapy adherence from the inertial data collected at home	AI-driven physiotherapy diminishes drop-out, the commonest causes of less effect of an exercise program. The concept lowers the threshold for good compliance and for study participation for patients in primary healthcare	[19]	Adherence tracking with smart watches for shoulder physiotherapy in rotator cuff pathology. Smart gadgets can be integrated into protocols of, e.g., longitudinal cohort studies in physiotherapy
	Clinic effectiveness	ML can help improve both clinic management per se (logistics, economy, human resources, admissions, etc.) and adherence to evidence-based treatment program and other routines, which increase the efficacy of applied treatments	Valid clinical or pharmacogenetic predictors of response are needed to tailor specific algorithms to individual patients	Algorithm research has evolved as a new branch of clinical research that evaluates the clinical and economic impact of algorithm-guided treatment in primary and psychiatric care of patients with depressive disorders	[69]	Treatment algorithms and collaborative-care systems are systematic treatment approaches that are designed to improve outcomes by enhancing the quality of care; e.g., such systems for treatment of depression
	Operational tasks	ML will boost primary healthcare management, with effects on, e.g., planning, waiting lists, scheduling, length of stays, workload, quality of care, etc.	E.g., a simulated annealing algorithm to optimize the patient admission sequence towards minimizing the total completion and total waiting of patients	Operational tasks in primary healthcare which can be augmented with ML include workflow, diagnostic imaging when expected in primary setting, personnel staffing and scheduling, process assessments, organizational case studies, and patient admission and referral routines	[70]	The example ML model have demonstrated to be effective in the evaluation of diagnostic imaging workflows
	Risk prediction	A tangible ML advantage for a primary healthcare clinic is naturally risk prediction	Random forest on QRISK3 risk prediction algorithms	Encapsulated a long list of risk factors and has identified potential new risk factors	[71]	Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease
	Task automation	Digital AI-embedded primary healthcare will enable thorough task automation	Achieving automation in the task delegation process can use several ML methods; neural networks, evolutionary algorithms, or swarm intelligence algorithms	ML can positively and negatively affect job resources (autonomy/control, skill use, job feedback, relational aspects) and job demands (e.g., performance monitoring)	[72]	Work design improvement – with consequences for employee well-being, safety, and performance

(continued)

Table 2 (continued)

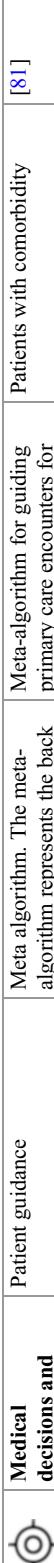
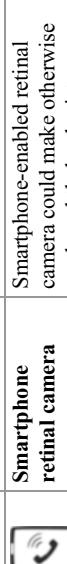
	Area	Action	Machine learning algorithm	Explanation	Example	Reference
	Individual treatment	Treatment based on a person's specific setup of genes, proteins, environment, medical history, and other conditions	Error-bound for estimating individual treatment effect, and creation of learning CNN algorithms	Family of algorithms for predicting individual treatment effect (ITE) from observational data, under the assumption known as strong ignorability	The bound relates ITE estimation to the classic machine learning problem of learning from samples, along with methods for measuring distributional distances from samples	[73]
	Precision medicine	Precision medicine (PM) for a long range of conditions, will be become omnipresent in primary healthcare	Key points to be addressed for a reproducible treatment algorithm: technology, thresholds, quality controls, validation of molecular alterations identified, interpretations of the results, molecular alteration/target relation, prioritization	Also called “personalised medicine,” is defined by the National Cancer Institute as “a form of medicine that uses information about a person’s genes, proteins, and environment to prevent, diagnose, and treat disease”	Treatment algorithms based on tumor molecular profiling in oncology	[74]
	Clinical trials	Design and recruitment of clinical trials, with many new possibilities as a result of AI, also interacting to closer with primary healthcare and the electronic health records and other databases	Both supervised and semi-supervised learning, cluster analysis, reinforcement learning (RL), in which the algorithm learns from previous experience and adjusts its behavior in response to what it has learned and causal inference	Machine learning will provide important insights into (1) ongoing clinical trials for non-COVID-19 drugs, (2) clinical trials for repurposing drugs to treat COVID-19, and (3) clinical trials for new drugs to treat COVID-19	Clinical trials have adapted significantly to new conditions during the SARS-CoV-2 pandemic – many times with the help of machine learning	[75, 76]
	Drug development	Deep drug development, individual precision medicine will move closer to the AI-rich future primary healthcare	A plethora of deep learning algorithms have been implemented in several drug discovery processes. ML is active on several levels; with DL, meta learning, general adversarial networks (GAN) reinforcement learning, support vector machines, symbolic learning, and convolutional neural networks (CNN)	Active in stages such as: (1) peptide synthesis, (2) dosage/delivery, (3) active/inactive ligand classification, (4) protein folding prediction, (5) structure based virtual screening, (6) drug repurposing (QSAR), (7) biomarker discovery and preclinical development, (8) bioactivity prediction,	Used in, e.g., peptide synthesis, structure-based virtual screening, ligand-based virtual screening, toxicity prediction, drug monitoring and release, pharmacophore modelling, quantitative structure–activity relationship, drug repositioning, polypharmacology, and physicochemical activity	[77]

			(9) molecular pathway identification, (10) mode-action prediction and ADMET, and (11) primary/secondary drug screening	The same authors have also shown how several other values, e.g., blood sugar, temperature, and HbA1c can be detected with this methodology; see their publication list [40]	[46, 78]
	Transdermal optical imaging	Smartphones scanning faces with the camera, and can give accurate blood pressure, pulse, and other vital figures	Advanced machine learning algorithm to create computational models that predict reference systolic, diastolic, and pulse pressure from facial blood flow data	With this technique, all essential vital sign can be given to the patient or doctor anywhere	
	Monitoring	Continuous monitoring of diseases moving out of hospitals and to the patient's homes and also closer to the primary healthcare	<i>ICU</i> : e.g., trend detection and curve fitting (e.g., the change in the average heart rate between the current and the previous minute is more than a specified value). Also, multivariate statistical methods. Various machine learning approaches were applied in this field. Deep neural networks with recurrent architecture are effective in speech recognition	Alarm algorithms were previously seen only in critical care monitoring. The alarms of medical devices are a matter of concern in critical and perioperative care with a high frequency of false alarms. Smartphones are now also capable monitoring tools allowing for continuous monitoring of many diseases reserved for the ICU	In the <i>ICU</i> , there is a new for more relevant alarms, with new, e.g., alarm algorithms using statistical approaches. Smartphones are increasingly used to monitor everything from diabetes, bipolar disorder, vital signs, etc., and start to mimic many of the tools, which were previously reserved for the ICU
	Symptom checker	Tools that use computer algorithms to help patients with self-diagnosis or self-triage	Eg., the DocBot clinical decision support algorithm	Symptom checkers (23 publicly available, and free checkers tested) have yet deficits in both triage and diagnosis. Triage advice from symptom checkers is generally risk averse, encouraging users to seek care for conditions where self-care is reasonable	[16]

(continued)

Table 2 (continued)

	Area	Action	Machine learning algorithm	Explanation	Example	Reference
	Dermatology	Assortment, triaging, guidance, and auto-diagnosis dermatological photos	Several common types of machine learning approaches used in dermatology: convolutional neural network (CNN), natural language processing (NLP), support vector machine, and random forest. Notably, there are many other possible machine learning approaches	Recent advancements in access to large datasets (e.g., electronic medical records, image databases, omics), faster computing, and cheaper data storage have encouraged the development of ML algorithms with human-like intelligence in dermatology	Potential to improve the dermatologist's practice from diagnosis to personalized treatment and speed up the connection with primary healthcare	[18]
	Pathology	Analysis of pathology slides, moving the speciality close to primary healthcare	Pathology image analysis using segmentation deep learning algorithms with convolutional neural networks	Full image scanning techniques and visualization software, whole slide imaging (WSI) has become routine	Acceleration of clinical diagnosis from pathology images and automating image analysis efficiently and accurately	[79]
	Medical photos	To categorize, diagnose, triage, and evaluate a wide range of medical imagery. Potential basis of computer vision systems	Multiple instance learning (MIL)	Suitable for image processing applications and based on a mixed integer nonlinear optimization problem	The algorithm has been preliminarily applied to a set of color images, with the aim to identify images containing some specific color pattern, and successively to a medical dataset, containing photos of melanoma and common nevi	[79]
	Radiology assistant	Virtual assistant in radiology for both radiologists and GPs. E.g., chest X-ray evaluation	An application using a convolutional neural network (CNN) model to operate as a predictive model for diagnosis support as a virtual assistant	This model of artificial intelligence (AI) provides seamless integration with the preexisting workflow without disrupting the regular procedure of the physicians	ToraxIA: Virtual assistant for radiologists based on deep learning from chest X-ray	[80]

	Medical decisions and guidance Patient guidance	<p>Meta-algorithm. The meta-algorithm represents the backbone of the multimorbidity guideline of the German College of General Practitioners and Family Physicians</p>	<p>Meta-algorithm for guiding primary care encounters for patients with evidence-based and case-based guideline development methodology</p>	<p>Patients with comorbidity [81]</p>
	Smartphone retinal camera 	<p>Smartphone-enabled retinal camera could make otherwise costly ophthalmologist technology more accessible</p>	<p>Retinal image analysis is a challenging problem due to the precise quantification required and the huge numbers of images produced in screening programs</p>	<p>Brain-inspired algorithms for automated retinal image analysis, recently developed for the RetinaCheck project. Many of these algorithms outperform state-of-the-art techniques. With deep residual networks diabetes can be diagnosed directly from retinal images, without using any blood glucose information</p>

Precision Medicine and Frontiers for AI

Precision medicine stems from identifying new patterns within the genetics of an individual and tailoring treatments to their disease. It is based on the premise that under all the identical genetic information, individuals have unique differences that distinguish them. Machine learning and AI are facilitating this process allowing tailored therapies through pharmacogenomics that can reduce issues with medication side effects and treatment failures [21, 22]. Emerging frontiers for AI in medicine are already entering primary care with the spread of apps via “passive diffusion.”

Legal and Regulatory Aspects

Social and legal considerations must be considered where medicine borders artificial intelligence, together with regulatory and political hindrances that surround aspects relating to who should own patient data [9, 23]. This will require public, policy-maker, and regulator engagement before AI could be accepted in the public eye for primary care. The speed at which newer and emerging technologies are incorporating AI into patient diagnostics using the patient’s data also present their own ethical challenges for primary healthcare. Questions such as who is to blame if an AI system misclassifies a disease fail to alert a physician or fail to follow specific instructions due to a bug or cyberthreat and become important discussions for the policy-maker, the family physician, and the patient. Another, perhaps more intrinsic, problem arises if patients do not perform self-tests in the proper ways, people screw the system by sending other’s people’s data or data get lost by a hack, etc.

Privacy Concerns

Additionally, privacy preservation when AI and the data used to build it is out of the owner’s control and into the hands of industrial giants like Google, Microsoft, and Facebook. Such corporations want access to patient data and argue in favor of utilizing the data to train AI algorithms in order to benefit

and improve patient healthcare. However, some patients feel that their data should not be used by these companies who do not incentivize patients or compensate them from the advertising revenue they generate from the use of such patient data. This adds a layer of scrutiny regarding safeguarding concerns, data ownership, and data security. The primary care physician and the patient themselves may not benefit from the applications being created by these platforms building AI systems while unethically compromising patient privacy.

Patient Safety

The above point raises a case for the consideration of patient safety and hence reducing the risk to patients especially the vulnerable with safeguarding concerns who attend primary care. Machine learning algorithms for prediction of risk and risk mitigation systems could also benefit primary care physicians and their patients. Patients will need education about risk and risk mitigation practices, and AI platforms analyzing the knowledge about a particular risk-factor determinant can link these patients to the right knowledge base.

Medical Imaging Diagnostics and Radiology

AI for medical diagnostics has also come a long way since developments, and arguments were first made for its use in medical image pattern recognition. Clinicians and radiologists are able to utilize machine learning for medical image diagnostics in secondary care. So, its use would be especially practical in primary healthcare systems where radiology support is usually limited. However, there is also the innate sense of superconservatism within medicine at various technology readiness levels. Technological innovation and its implementation post-regulation can take more than a decade to reach the clinical environment as this usually requires extensive clinical trials. On the other hand, wearable technologies were introduced over a very short period of time

and have now become part of the lives of many patients. Wearables with AI capabilities are needed to analyze and alert primary care clinicians. Diagnostic laboratory investigations usually require interpretation for the patients to understand what their tests mean. This can occur offline and through mHealth systems.

Explainable methods in medical imaging and diagnostics are also benefiting from AI. There are newer primary care systems such as those proposed by Prof. Lord Darzi's GP-led Polyclinics, which were set up in 2008 and had self-contained services with clinical imaging facilities such as X-ray, CT, and ultrasound [24]. These would benefit from augmentations from AI-based infrastructures that support specialist-guided auto-interpretation achieved through supervised learning approaches.

Medical Informatics and Clinical Decision Support

Evidence-based practice is promoted across the globe for medical informatics and evidence-backed treatment. The difficulty is that time pressures preclude effective appraisal of the literature. Medical informatics and augmented evidence-based medicine in primary care are becoming more feasible using artificial intelligence platforms that are reducing the time needed to appraise the evidence. AI-powered platforms that can aid in systematic literature searches and appraisals are now available or are being proposed, and this can support evidence-based information discovery.

As described by Tversky, clinical decision-making utilizes cognitive heuristics for recall or to understand knowledge, and GPs may usually generate a range of likely differential diagnoses that often require evidence-based methods to rule in or out [5, 25]. There is seemingly an expectation bias, which can lead to the dismissal of important information that is deemed less relevant. Areas identified to be affected by this methodology of diagnostics are geographically variable. For instance, in the UK and many other parts of the world, earlier diagnosis of cancer continues to remain a challenge, and this might reflect the use of heuristics. Sheringham et al. conducted a factorial experiment that showed that

GPs are not more likely to initiate cancer investigations for individuals with higher-risk symptoms and, also, do not investigate everyone with the same symptoms equally [5, 26]. There was a potential 42% rate of omission to seek further information from their patients perhaps because of underlying cognitive biases.

AI could assist GPs in recognizing and overcoming cognitive biases, but only if the system has been trained on the correct data and from the appropriate population. An AI diagnostic tool developed using clinical information from the UK might not provide the correct clinical guidance to a primary care clinician in India. Moreover, the AI outputs need to support GP decision-making where, for example, precise diagnostic labels are less important than deciding on an appropriate course of action.

In enhancing a GP's diagnostic abilities, it is also important to appreciate that AI systems do not get tired or irritable, which can affect a GP's ability to communicate and work effectively. Moreover, AI might have a key role in assessing and improving GP communication skills [27]. It must clearly be mentioned that there is a rapid improvement of AI systems over time by continuous self-learning. These systems are not static. AI systems of just a few years ago are considered primitive nowadays.

Patient's Perspective

Patients prefer rapid and easy access to specialist services through primary care, extended consultation times, a clinician (usually their GP) to be available to support their queries at a moment's notice and control over their data. Some of these aspects, such as their own GP being available around the clock, are impractical. However, most patient-centered systems in some developed economies deliver this to a large extent as out-of-hours services, but this can be a challenge in other emerging healthcare economies. Some of these challenges can be overcome using AI. In fact, chatbots and robotic process automation systems are being used to support out-of-hours practice and

triaging. Given the growing globalization of healthcare, this network of primary care support will continue to evolve; however, challenges still exist in this regard [28, 29].

Software bots perform precise, preset workflows to reduce the potential for errors and the associated costs, and they can do this around the clock, without a drop-off in efficiency or quality [29]. Systems like robotic process automation have the potential not only to cut operating costs by streamlining workflows but also to ensure greater compliance, which in turn feeds through into higher levels of patient satisfaction.

Gender Aspects

Gender differences in disease usually affect the management of patients, and AI can support evidence-based guidance identifying where differences could influence treatment outcome and identify at-risk diagnoses to enhance primary care treatment of chronic diseases [30]. Initially, the design of these algorithms did not take these differences into consideration, but this is changing, and chapter two of this handbook demonstrates how AI is being implemented to study gender differences in areas such as reproductive medicine.

Point-of-Care Dermatology and Ophthalmology

By using machine learning platforms as point-of-care diagnostic tools to diagnose skin lesions, patients are able to take photos of skin lesions that can be classified using artificial intelligence and triaged for dermatological input [18, 31]. Primary care physicians with specialist interest in dermatology can leverage these platforms to manage skin lesions, together with specialist dermatologists. Dermatological diagnosis before a patient attends a clinic has its limitations, but also offers possibilities with AI augmentation to rapid diagnostics and referral for early surgical intervention [31]. AI has also shown potential in interpreting many different types of image data including retinal scans (discussed below), radiographs, and ultrasound [32, 33]. Many of these

images can be captured with relatively inexpensive and widely available equipment.

Public Health Aspects on Primary Healthcare

In the context of public health management of epidemics and for contact tracing, AI systems were developed to support viral epidemic and pandemic tracking for flu, human immunodeficiency virus (HIV), and now SARS-COV-2. Monitoring of HIV and other sexually transmitted infections requires a concerted multidisciplinary effort by virologists with primary care physicians who may have specialist interest in public health or sexually transmitted infectious diseases, and the cooperation of patients. AI tracking systems designed to track symptomatic individuals were developed during the SARS-COV-2 pandemic in most countries including the USA and South Korea. This adds an additional layer of privacy concerns that makes it difficult for regulators and policy makers. Although rules seemed relaxed during SARS-COV-2 pandemic.

SARS-Cov-2 has also amplified the increasingly important role of primary care especially in managing mental health, chronic disease management, and SARS-COV-2-related issues. From a health system perspective, primary care is essential to the Triple AIM of healthcare reform [34]:

1. Population health
2. Experience of care
3. Per capita cost

An AI system to support or improve these aims is desired.

AI for General Practice Management

AI can also automate repetitive clerical tasks. Eligibility checks, insurance claims, prior authorizations, appointment reminders, billing, data reporting, and analytics can all now be automated using AI, and some companies have developed AI-powered category auditors to help optimize coding for quality payment programs.

Endocrinology and Diabetes

Endocrinological diseases can be highly complex, and specialist knowledge is required to manage patients presenting with these conditions. Patients usually require adjustments to medications and routine biochemical investigations to check their thyroid, adrenal, and hypothalamic function [35]. For diabetics, blood sugar and HbA1c check-ups are required to optimize diabetic management and identify poor blood glucose control, and diabetic ophthalmic complications from micro- and macro-angiopathic changes [32]. Current implantable platforms for AI-based continuous monitoring of diabetes have been designed to support primary care management of diabetic disorders [36, 37].

Selection of the right combination of drugs carries combinatorial complexities with over 11 factorial possibilities for optimal drug class combinations with added pharmacogenomic complexity [21]. This may benefit from artificial intelligence systems that take into account poly-medication that the patient is already taking, clinical information, investigations, follow-up, patients' life style choices, and risk assessment.

The federal Food and Drug Administration approved IDx, a convolutional neural network-centered diagnostic system that uses AI to detect diabetic retinopathy [32]. IDx does not require a specialist to interpret the images or results, making it the first such system cleared for use by the FDA. Physicians and their care teams, even those not normally involved in eye care, can use the technology to screen their patients for the condition during routine office visits.

Cardiovascular and Respiratory Management

Cardiac

Wearable devices like the Apple watch that can remind patients to take their cardiac medications and to monitor irregular heart rhythms can send data to the general practitioner and ensure timely treatment or early warning about an impending cardiac arrest. Automated implantable

defibrillating systems have algorithms that can detect an arrest to activate a defibrillator and that can predict electrical storms [38]. The management of these devices if they malfunction is not routinely in the control of the primary care physician, but patients are expected to troubleshoot any issue independently to inform the specialist. Assistive AI technologies could support the primary care physician in troubleshooting a device malfunction. The physician can be alerted to provide support and could also prompt a direct referral to the specialist team or alert the specialist-interested primary care physician and keep them in the communication loop. AI systems for remote diagnostics including ultrasound scans for a cardiologist's remote ECG interpretation, remote ultrasound, and remote diagnostic reporting will all become beneficial for the patient and primary care physician [39]. Cardiogenic and neuro-syncopal events usually need to be differentiated from each other, and these are areas that diagnostic machine learning could also support.

Hypertension

Hypertension management can require regular and ambulatory monitoring. Its management utilizes six main classes of drug types and can also benefit from pharmacogenomics and drug choice prediction, and hypertensive complication management using artificial intelligence [21]. Follow-up systems would benefit from continuous machine-learning-based systems monitoring with smartphone apps, transdermal optical imaging to monitor yearly fluctuations in blood pressure as a determinant to cardiovascular risk [40].

Respiratory

Several chronic respiratory disorders require medical and rehabilitation support such as chronic obstructive pulmonary disorders and asthma are benefiting from artificial intelligence [41]. Other respiratory pathologies such as cystic fibrosis and other restrictive pathologies, together with those that are associated with neuromuscular junction and neurodegenerative disorders leading to

respiratory complications, are areas that will see increasing support from artificial intelligence. Machine learning models deployed through m-Health platforms and smartphone apps are helping to support patient rehabilitation and management of their disease as well as decision support for the primary care physicians managing complex respiratory patients [42].

Chronic Neurological and Neuropsychiatric Disease Monitoring

Neurological diseases like Parkinson's disease and various other diseases can affect gait, fine motor dexterity, and other systems [43, 44]. Others such as epilepsy and associated neuropsychiatric conditions like depression, schizophrenia, eating disorders are also managed by primary care physicians, etc. [44–46]. Artificial intelligence systems through wearables have been designed to support the diagnoses of diseases through assessment of subtle gait disturbances [43]. Seizure monitoring systems are available that either enable alerts for the physician to be made aware of changes in a patient's seizure activity or for predictors of an impending seizure to be differentiated from black-out events. Other conditions such as alcohol and drug dependence are also utilizing AI to support rehabilitation. Autism is also an area that AI will continue to develop to support. Cerebrovascular disease support for rehabilitation, diagnostics, and at-risk prediction of strokes have all benefitted from machine learning models. Relevant chapters in this handbook have discussed these areas in depth, and the reader is referred to the AIM in neurology, stroke, and neurodegenerative disorders sections for more insights.

Obstetrics, Pregnancy, and Pediatrics

Pregnancy monitoring and the monitoring of physiological parameters for the developing baby's fetal heartrate interpretation, dating scans, Down's syndrome ultrasound prediction, supporting detection for neonatal infections, and identifying

developmental abnormalities in children. Others are developing AI applications for screening for adverse perinatal outcomes [47]. Such abnormalities including speech deficits are usually incorporated in the screening undertaken by health visitors and some primary care physicians with specialist interest in obstetrics and pediatrics. Unsurprisingly, these areas stated above will also be benefitting from artificial Intelligence, and machine learning methods have been used to support medication choices, etc. [48, 49].

Oncology

Oncology remains a highly complex and challenging disease carrying a large global financial burden where diagnostics and treatment choice is limited. As such early involvement of the specialist oncologist and surgeon is necessary for diagnostics, surgical management, neoadjuvant chemoradiotherapy, and adjuvant therapy.

AI-powered algorithms for diagnosing disease are outperforming physicians in detecting skin cancer, breast cancer, colorectal cancer, and brain cancers [32]. Numerous tools, such as IDx-DR, Aysa, and Tencent that can reduce the need for unnecessary referrals, increase continuity with patients and enhance mastery for primary care physicians [32, 50].

Newer AI platforms combining symptom checkers with clinical pattern recognition algorithms could facilitate remote oncological diagnostics, which will monitor clinical features and identify patients at risk of cancers for primary care physician referral, direct specialist referral, or predict cancer risk for early detection and management. This will be facilitated by wearable technologies, AI-powered implantable devices, and point-of-care diagnostic set-ups. Further developments of these technologies will be crucial for specialties such as neuro-oncology in brain tumor detection for at-risk groups, etc. AI-powered radiogenomic diagnostics will also enable and facilitate early screening and pharmacogenomic interventions leading to early referral processes and

selections for immediate appropriate medical and surgical therapies. It would also facilitate engagement with specialist teams of oncologists allowing earlier information flow between them and the primary care physician.

Many other primary care-related conditions from other body systems and subspecialties will continue to benefit from artificial intelligence as summarized in Tables 1 and 2 together with the type of machine learning models that have been developed to solve the issues. Table 1 also provides a summary overview (June 2021) of some of the companies that are helping to innovate within this space.

Conclusion

In conclusion, primary healthcare is potentially where AI in medicine will be most visible to most people. Developments will include AI empowering independence, patient ownership, and responsibility for their data with the general practitioner being able to access this data easily.

Primary healthcare will not necessarily mean a physical clinic that we are used to – it can take place online, in the street, in a clinic, on the go, in the patient's pocket or in the body, over the phone or webcam, or at home supported by machine learning algorithms to secure the interaction.

A suite of AI tools will be developed for the patient and their primary healthcare network, where a plethora of apps, clinical decision support, transparent and two-way educative electronic health records. This may utilize a cloud-based infrastructure constructed in such a way that the patient has independent access to their data under the guiding principles of what is legal and within that particular country's ethico-legal remit.

Moreover, the shift of balance of care may edge closer towards the primary care physician as we are seeing in some developed economies. Some proponents seem to believe that the Scandinavian model of pluripotent healthcare centers, at least in remote areas, may be easier to establish with embedded AI. However, other systems are bound to evolve to meet the challenge of incorporating AI into their

primary healthcare systems. This will also bring new opportunities for emerging economies.

The traditional limitations of homes, hospitals, primary clinics, and cyberspace will be increasingly blurred. The locations of treatment and also that of diagnostics will evolve. Most management methods may be moved to the home setting to meet the demand for excellent care of patients with special needs, the elderly, and patients with conditions after stroke or with neurodegenerative diseases. The use of AI will enable advanced medical monitoring at home, even in critical cases with available multidisciplinary primary care physician support for managing complex cases. A benefit of primary healthcare is its cross-disciplinary nature. This chapter has also demonstrated how AI can be used to triage the patient, far in advance and far away. Multiple medical specialties can be tied into this chapter, and in some ways this chapter concludes the impact of AI in medicine as a whole, but readers are referred to relevant chapters for added inspiration and in-depth overview. We have also highlighted some of the major trends, recognizing that a summary of such a vast field is a real challenge. Primary healthcare will be a key outlet for effective design and delivery of AI in medicine, since it has a central role of coordinating and involving specialties.

Cross-References

- ▶ [AIM in Alcohol and Drug Dependence](#)
- ▶ [AIM in Allergy](#)
- ▶ [AIM and Brain Tumors](#)
- ▶ [AIM and Business Models of Healthcare](#)
- ▶ [AIM and Causality for Precision and Value-Based Healthcare](#)
- ▶ [AIM in Clinical Neurophysiology and Electroencephalography \(EEG\)](#)
- ▶ [AIM in Depression and Anxiety](#)
- ▶ [AIM in Dermatology](#)
- ▶ [AIM and Diabetes](#)
- ▶ [AIM in Electronic Health Records \(EHRs\)](#)
- ▶ [AIM in Endocrinology](#)

- AIM and Explainable Methods in Medical Imaging and Diagnostics
- AIM in Genomic Basis of Medicine: Applications
- AIM and Gender Aspects
- AIM in Genomics
- AIM in Health Blogs
- AIM and mHealth, Smartphones and Apps
- AIM in Medical Disorders in Pregnancy
- AIM in Medical Informatics
- AIM in Neurology
- AIM in Nursing Practice
- AIM in Osteoporosis
- AIM in Oncology
- AIM and Patient Safety
- AIM and the Patient's Perspective
- AIM in Rehabilitation
- AIM in Respiratory Disorders
- AIM in Rheumatology
- AIM and Transdermal Optical Imaging
- AIM in Wearable and Implantable Computing
- AI in Forensic Medicine for the Practicing Doctor
- Artificial Intelligence in Medicine (AIM) for Cardiac Arrest
- Artificial Intelligence in Medicine (AIM) in Cardiovascular Disorders
- Artificial Intelligence in Clinical Toxicology
- Artificial Intelligence in Epidemiology
- Artificial Intelligence in Evidence-Based Medicine
- Artificial Intelligence in Forensic Medicine
- Artificial Intelligence for Medical Decisions
- Artificial Intelligence in Ophthalmology
- Artificial Intelligence in Acute Ischemic Stroke
- Artificial Intelligence in Medicine in Anemia
- Artificial Intelligence and Hypertension Management
- Artificial Intelligence for Physiotherapy and Rehabilitation
- Artificial Intelligence in Medicine and Privacy Preservation
- Artificial Intelligence for Autism Spectrum Disorders
- Artificial Intelligence in Pediatrics
- Artificial Intelligence in Public Health
- Artificial Intelligence in Telemedicine
- Artificial Intelligence in Schizophrenia
- Emergence of Deep Machine Learning in Medicine

- Machine Learning and Electronic Noses for Medical Diagnostics
- Meta Learning and the AI Learning Process
- The New Frontiers of AI in Medicine

References

1. Alami H, Rivard L, Lehoux P, Hoffman SJ, Cadeddu SBM, Savoldelli M, Samri MA, Ag Ahmed MA, Fleet R, Fortin JP. Artificial intelligence in health care: laying the foundation for responsible, sustainable, and inclusive innovation in low- and middle-income countries. *Glob Health.* 2020;16(1):52.
2. Bloom D, et al. The Global Economic Burden of Non-Communicable Diseases. A report by the World Economic Forum and the Harvard School of Public Health, September 2011. 2011.
3. WHO. Global Health Workforce alliance and World Health Organization. A universal truth: no health without a workforce. Available online: https://www.who.int/workforcealliance/knowledge/resources/GHWA-a_universal_truth_reportpdf?ua=1. 2013.
4. Li L. Artificial intelligence and diagnosis in general practice. *Br J Gen Pract.* 2019;69(686):430.
5. Summerton N, Cansdale M. Artificial intelligence and diagnosis in general practice. *Br J Gen Pract.* 2019;69(684):324–5.
6. Rappoport N, Shamir, R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark [published correction appears in Nucleic Acids Res. 2019 Jan 25;47(2):1044]. *Nucleic Acids Res.* 2018;46(20): 10546–10562.
7. Imison C, Curry, N, Holder, H, Castle-Clarke, S, Nimmons, D, Appleby, J, Thorlby, R and Lombardo, S. Shifting the balance of care: great expectations. Research report Nuffield Trust. 2017.
8. Sinsky C, Colligan L, Li L, Prgomet M, Reynolds S, Goeders L, Westbrook J, Tutty M, Blike G. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Ann Intern Med.* 2016;165(11):753–60. Epub 2016 Sep 6. PMID: 27595430.
9. Liaw S, Liyanage H, Kuziemsky C, Terry AL, Schreiber R, Jonnagaddala J, de Lusignan S. Ethical use of electronic health record data and artificial intelligence: recommendations of the primary care informatics Working Group of the International Medical Informatics Association. *Yearb Med Inform.* 2020;29(1):51–7.
10. Liaw W, Kakadiaris IA. Primary care artificial intelligence: a branch hiding in plain sight. *Ann Fam Med.* 2020;18(3):194–5.
11. Liyanage H, Liaw ST, Jonnagaddala J, Schreiber R, Kuziemsky C, Terry AL, de Lusignan S. Artificial intelligence in primary health care: perceptions, issues, and challenges. *Yearb Med Inform.* 2019;28(1):41–6. <https://doi.org/10.1055/s-0039-1677901>.

12. NHS. The Topol review: the Topol Review — NHS Health Education England. NHS Health Education England Retrieved 2020-03-11. 2018.
13. Topol E. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25(1):44–56.
14. Peiffer-Smadja N, Rawson TM, Ahmad R, Buchard A, Georgiou P, Lescure FX, Birgand G, Holmes AH. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. *Clin Microbiol Infect.* 2020;26(5):584–95.
15. Ghatnekar S, Faletsky A, Nambudiri VE. Digital scribe utility and barriers to implementation in clinical practice: a scoping review. *Health Technol (Berl).* 2021;11:1–7.
16. Semigran H, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. *BMJ.* 2015;351:h3480.
17. Kroenke K, Jackson JL. Outcome in general medical patients presenting with common symptoms: a prospective study with a 2-week and a 3-month follow-up. *Fam Pract.* 1998;15(5):398–403.
18. Chan S, Reddy V, Myers B, et al. Machine learning in dermatology: current applications, opportunities, and limitations. *Dermatol Ther (Heidelb).* 2020;10:365–86.
19. Burns D, Razmjou H, Shaw J, Richards R, McLachlin S, Hardisty M, Henry P, Whyne C. Adherence tracking with smart watches for shoulder physiotherapy in rotator cuff pathology: protocol for a Longitudinal Cohort Study. *JMIR Res Protoc.* 2020;9(7):e17841.
20. Myburgh H, Jose S, Swanepoel D, Laurent C. Towards low cost automated smartphone- and cloud-based otitis media diagnosis. *Biomed Signal Process Cont.* 2018;39:34–52.
21. Silva PJ, Jacobs D, Kriak J, Abu-Baker A, Udeani G, Neal G, Ramos K. Implementation of pharmacogenomics and artificial intelligence tools for chronic disease management in primary care setting. *J Pers Med.* 2021;11:443.
22. Primorac D, et al. Pharmacogenomics at the center of precision medicine: challenges and perspective in an era of Big Data. *Pharmacogenomics.* 2020;21(2):141–56.
23. Guan J. Artificial intelligence in healthcare and medicine: promises, ethical challenges and governance. *Chin Med Sci J.* 2019;34(2):76–83.
24. Darzi A. High quality care for all – NHS next stage review final report Department of Health. 2008.
25. Tversky A, Kahneman D. Judgement under uncertainty: heuristics and biases. *Science.* 1974;185(4157):1124–31.
26. Sheringham J, Sequeira R, Myles J, et al. Variations in GPs' decisions to investigate suspected lung cancer: a factorial experiment using multimedia vignettes. *BMJ Qual Saf.* 2017;26(6):449–59.
27. Ryan P, Luz S, Albert P, Vogel C, Normand C, Elwyn G, et al. Using artificial intelligence to assess clinicians' communication skills. *BMJ.* 2019;364:l161.
28. Miles O. Acceptability of chatbot versus General Practitioner consultations for healthcare conditions varying in terms of perceived stigma and severity (Preprint). Qeios. 2020; <https://doi.org/10.32388/BK7M49>.
29. Willis M, Duckworth P, Coulter A, Meyer ET, Osborne M. The future of health care: protocol for measuring the potential of task automation grounded in the National Health Service Primary Care System. *JMIR Res Protoc.* 2019;8(4):e11232.
30. Cirillo D, Catuara-Solarz S, Morey C, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ Digit Med.* 2020;3:81.
31. Zakhem G, Motosko CC, Ho RS. How should artificial intelligence screen for skin cancer and deliver diagnostic predictions to patients? *JAMA Dermatol.* 2018;154(12):1383–4.
32. Abràmoff M, Lou Y, Erginay A, et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Invest Ophthalmol Vis Sci.* 2016;57:5200–6.
33. Ting D, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, Tan GSW, Schmetterer L, Keane P, Wong TY. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol.* 2019;103(2):167–75.
34. Verma A, Bhatia S. A policy framework for health systems to promote triple aim innovation. *Healthc Pap.* 2016;15(3):9–23.
35. Gubbi S, Hamet P, Tremblay J, Koch CA, Hannah-Shmouni F. Artificial intelligence and machine learning in endocrinology and metabolism: the Dawn of a New Era. *Front Endocrinol (Lausanne).* 2019;10:185. Published 2019 Mar 28. <https://doi.org/10.3389/fendo.2019.00185>
36. Vettoretti M, Cappon G, Facchinetto A, Sparacino G. Advanced diabetes management using artificial intelligence and continuous glucose monitoring sensors. *Sensors (Basel).* 2020;20(14):3870. Published 2020 Jul 10
37. Van Doorn WPTM, Foreman YD, Schaper NC, Savelberg HHCM, Koster A, et al. Machine learning-based glucose prediction with use of continuous glucose and physical activity monitoring data: the Maastricht Study. *PLoS One.* 2021;16(6):e0253125.
38. Shakibfar S, Krause O, Lund-Andersen C, Aranda A, Moll J, Andersen TO, et al. Predicting electrical storms by remote monitoring of implantable cardioverter-defibrillator patients using machine learning. *EP Europace.* 2018;21(2):268–74.
39. Ribeiro AH, Ribeiro MH, Paixão GMM, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat Commun.* 2020;11:1760.
40. Luo H, et al. Smartphone-based blood pressure measurement using transdermal optical imaging technology, circulation. *Cardiovasc Imaging.* 2019;12(8):e008857. <https://doi.org/10.1161/CIRCIMAGING.119.008857>. Epub 2019 Aug 6. PMID: 31382766

41. Kaplan A, Cao H, FitzGerald JM, Iannotti N, Yang E, Kocks JWH, et al. Artificial intelligence/machine learning in respiratory medicine and potential role in asthma and COPD diagnosis. *J Allergy Clin Immunol Pract.* 2021;9(6):2255–61.
42. Liyanage H, Liaw ST, Jonnagaddala J, et al. Artificial intelligence in primary health care: perceptions, issues, and challenges. *Yearb Med Inform.* 2019;28(1):41–6.
43. Pedersen M, Verspoor K, Jenkinson M, Law M, Abbott DF, Jackson GD. Artificial intelligence for clinical decision support in neurology. *Brain Commun.* 2020;2(2):fcaa096. Published 2020 Jul 9.
44. Cavedoni S, Chirico A, Pedroli E, Cipresso P, Riva G. Digital biomarkers for the early detection of mild cognitive impairment: artificial intelligence meets virtual reality. *Front Hum Neurosci.* 2020;14(245). Published 2020 Jul 24. <https://doi.org/10.3389/fnhum.2020.00245>
45. Raghavendra U, Acharya UR, Adeli H. Artificial intelligence techniques for automated diagnosis of neurological disorders. *Eur Neurol.* 2019;82(1–3):41–64.
46. Anna Z, et al. Smartphone as a monitoring tool for bipolar disorder: a systematic review including data analysis, machine learning algorithms and predictive modelling. *Int J Med Inform.* 2020;138:104131.
47. Feduniw S, Sys D, Kwiatkowski S, Kajdy A. Application of artificial intelligence in screening for adverse perinatal outcomes: a protocol for systematic review. *Medicine.* 2020;99(50):e23681.
48. Davidson L, Boland MR. Enabling pregnant women and their physicians to make informed medication decisions using artificial intelligence. *J Pharmacokinet Pharmacodyn.* 2020;47:305–18.
49. Iftikhar P, et al. Artificial intelligence: a new paradigm in obstetrics and gynecology research and clinical practice. *Cureus.* 2020;12(2):e7124. Published 2020 Feb 28.
50. Jones O, et al. Artificial intelligence techniques that may be applied to primary care data to facilitate earlier diagnosis of cancer: systematic review. *J Med Internet Res.* 2021;23(3):e23483.
51. Pak M, Kim S. A review of deep learning in image recognition. 2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT), 2017, p. 1–3.
52. Ker J, et al. Deep learning applications in medical image analysis. in *IEEE Access*, vol 6, p. 9375–9389, 2018.
53. Zame W, Bica I, Shen C, Curth A, Lee H-S, Bailey S, et al. Machine learning for clinical trials in the era of COVID-19. *Stat Biopharm Res.* 2020;12(4):506–17.
54. Bae K, et al. Pulmonary nodules: automated detection on CT images with morphologic matching algorithm—preliminary results. *Radiology.* 2005;236(1):286–93.
55. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal.* 2017;42:60–88.
56. Gunçar G, Kukar M, Notar M, et al. An application of machine learning to haematological diagnosis. *Sci Rep.* 2018;8:411.
57. Schubart J, Fowler CE, Donowitz GR, Connors AF Jr. Algorithm-based decision rules to safely reduce laboratory test ordering. *Stud Health Technol Inform.* 2001;84(Pt 1):523–7. PMID: 1160479.
58. Dalton L, Ballarin V, Brun M. Clustering algorithms: on learning, validation, performance, and applications to genomics. *Curr Genomics.* 2009;10(6):430–445. <https://doi.org/10.2174/138920209789177601>
59. Yang L, et al. Gut microbiota-based algorithms in the prediction of metachronous adenoma in colorectal cancer patients following surgery. *Front Microbiol.* 2020;11:1106.
60. Thomas A, Barriere S, Broseus L, et al. GECKO is a genetic algorithm to classify and explore high throughput sequencing data. *Commun Biol.* 2:222, 2019.
61. Webster P. Patient data in the cloud. *Lancet, Digital Health.* 2019;1(8):E391–2.
62. Haras C, et al. Patient data synchronization process in a continuity of care environment. *AMIA Annu Symp Proc.* 2005;2005:296–300.
63. Yu T. The design of electronic medical records system using Skip-gram algorithm. *Netw Model Anal Health Inform Bioinform.* 2021;10:7.
64. Google. GoogleCloud Platform. Online Accessed June 2021. <https://www.cloud.google.com/ai-platform/training/docs/algorithms>
65. Schrödle B, Held L. Spatio-temporal disease mapping using INLA. *Environmetrics.* 2011;22:725–34.
66. Delgadillo J. Machine learning: a primer for psychotherapy researchers. *Psychother Res.* 2021;31(1):1–4.
67. Oh K, Lee, D, Ko, B, Choi, H. A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation. 18th IEEE International Conference on Mobile Data Management (MDM), 2017, p. 371–375.
68. Nasehi S, Pourghassem H. Seizure detection algorithms based on analysis of EEG and ECG signals: a survey. *Neurophysiology.* 2012;44:174–86.
69. Adli M, et al. Algorithms and collaborative-care systems for depression: are they effective and why?: a systematic review. *Biol Psychiatry.* 2006;59(11):1029–38.
70. Granja C, Almada-Lobo B, Janela F, Seabra J, Mendes A. An optimization based on simulation approach to the patient admission scheduling problem using a linear programming algorithm. *J Biomed Inform.* 2014;52:427–37.
71. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ.* 2017;357:j2099.
72. Parker SK, Grote G. Automation, algorithms, and beyond: why work design matters more than ever in a digital world. *Appl Psychol Int Rev.* 2020;1–45. <https://doi.org/10.1111/apps.12241>

73. Shalit U, Johansson, F, Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. Proceedings of the 34th International Conference on Machine Learning. PMLR 70: 3076–3085, 2017.
74. Christophe Le Tourneau MK, Tsimberidou A-M, Bedard P, Pierron G, Callens C, Rouleau E, Vincent-Salomon A, Servant N, Alt M, Rouzier R, Paoletti X, Delattre O, Bièche I. Treatment algorithms based on tumor molecular profiling: the essence of precision medicine trials. *JNCI*. 2016;108(4):djh362.
75. Zame W, Bica I, Shen C, Curth A, et al. Machine learning for clinical trials in the era of COVID-19. *Stat Biopharm Res*. 2020;12(4):506–17.
76. van Ginneken B. Grand challenges. Available from: <https://grand-challenge.org/>
77. Gupta R, Srivastava D, Sahu M, et al. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Mol Divers*. 2021;25:1315–1360. <https://doi.org/10.1007/s11030-021-10217-3>
78. Imhoff MK, Kuhls S. Algorithms in critical care monitoring. *Anesth Analg*. 2006;102(5):1525–37.
79. Wang S, Yang DM, Rong R, Zhan X, Xiao G. Pathology image analysis using segmentation deep learning algorithms. *Am J Pathol*. 2019;189(9):1686–98.
80. Carnier M, Gavidia L, Severeyn E, La Cruz A. ToraxIA: virtual assistant for radiologists based on deep learning from chest x-ray. *Artificial Intelligence, Computer and Software Engineering Advances*. 2021;1326:49–63. Published 2021 Feb 15.
81. Muche-Borowski C, Lühmann D, Schäfer I, The Guideline Group of the German College of General Practice and Family Medicine (DEGAM), et al. Development of a meta-algorithm for guiding primary care encounters for patients with multimorbidity using evidence-based and case-based guideline development methodology. *BMJ Open*. 2017;7:e015478.
82. ter Haar Romeny BM, Bekkers EJ, Zhang J, et al. Brain-inspired algorithms for retinal image analysis. *Mach Vis Appl*. 2016;27:1117–35. <https://doi.org/10.1007/s00138-016-0771-9>.
83. Abbasi-Sureshjani S, Dashtbozorg B, ter Haar Romeny BM, Fleuret F. Exploratory study on direct prediction of diabetes using deep residual networks. In: Tavares J, Natal Jorge R, editors. *VipIMAGE 2017. ECCOMAS 2017, Lecture notes in computational vision and biomechanics*, vol. 27. Cham: Springer; 2018. https://doi.org/10.1007/978-3-319-68195-5_86.



Danny D. Meetoo and Bertha Ochieng

Contents

Chapter Introduction	744
Introducing the IRs	744
Brief History of AI	745
Made not Born: The Beginning	745
Definition	747
Subsets of AI	747
Supervised Machine Learning (SML)	748
Unsupervised Machine Learning (UML)	749
Reinforcement Machine Learning (RML)	749
A Case for ICU	749
Deep Machine Learning (DML)	751
Nursing	751
Nursing Practice	752
Societal Change	752
AI and Nursing Education	753
Enhancing Holistic Care	753

D. D. Meetoo: deceased.

D. D. Meetoo
School of Nursing, Midwifery and Social Work, University
of Salford, Greater Manchester, UK

B. Ochieng (✉)
Integrated Health and Social Care, Faculty of Health &
Life Sciences, De Montfort University, Leicester, UK
e-mail: bertha.ochieng@dmu.ac.uk

Robots to the Rescue	754
The Future of Robotics	754
Conclusion	754
References	755

Abstract

Global emergence and exponential growth of artificial intelligence (AI) is becoming seamlessly integrated in our lives. By definition, AI describes a set of advanced technologies enabling machines to perform highly complex tasks effectively otherwise requiring intelligence that can't be matched if human beings were to perform them. How this reflects on the human mind is certain to continue being an area of ongoing research. This said, science fiction books and the numerous AI-induced apocalyptic scenarios presented to us in films and television series depicting AI gone wrong have cultivated a sense of fear and apprehension. However, to date, the healthcare industry does not evoke Orwellian concerns. Instead, the impact of this technology is particularly evident in the healthcare sector where this innovation promises transformation of the landscape of nursing practice along with the process involved in the delivery of collaborative, compassionate, ethically sound, and evidence-based patient care. In so doing, AI will likely resolve the current problem of global staff shortage and the decline in funding growth driven by an aging baby boomer generation living with multiple complex chronic health conditions.

During this transformative time, nurses must reflect on how healthcare technologies can ensure the patient's experience of care is embedded with understanding and compassion. The loss of these essential characteristics may lead patients to feel that in an AI-driven world their rights become an after-thought in the relentless pursuit of efficiency. To address this, formal and informal educational programs need to be reviewed. Besides enhancing the educators' AI knowledge, it would be

logical to include AI-related professionals to create a truly collaborative approach to training nurses.

Keywords

Artificial intelligence · Nursing · Education · Robotics · Technological competence

Chapter Introduction

The implications of artificial intelligence (AI) as the fourth industrial revolution (IR) on healthcare cannot be discussed without making any reference to the previous and related innovations. To this end, this chapter will outline the emergence of the preceding three IRs and their impact on society. It will then proceed to examine some of the key moments in AI. After defining AI and explaining the subsets of AI, this chapter will discuss how AI in healthcare can further influence a seamless quality of nursing practice care.

Introducing the IRs

During eighteenth-century England, modernity was conceived by technology. The resulting innovations which started the industrial revolutions fostered modern democracy together with the foundation of modern economies. Furthermore, society was disrupted when new buildings replaced the old structures. Similarly, institutions, industry, and demography were all to change course so that very little about life in Britain could be said to have been untouched by the revolution.

Today then we are witnessing an era of a technological revolution predicted to fundamentally change our lives as never seen before. To avoid

any confusion, it is perhaps appropriate to clarify what is being meant by the term “revolution.” The Cambridge Dictionary, for example, defines the term industrial revolution as “[...] the period of time during which work began to be done more by machines in factories than by hand at home” [1]. To this end, there is no doubt that advances in science and technology (S&T) have had a positively profound effect on the development of global industrialisation [2].

At this juncture, it is also noteworthy that the word revolution itself, as in the American Revolution and the French Revolution, implies the notion of suddenness, lasting for a few years. The industrial revolution, on the other hand, was not a sudden historical event but one which, although fast-paced, lasted for many years. Such phenomenon has been described as revolutionary while occurring over a long period of time. The scientific revolution of the sixteenth and seventeenth centuries, for example, is a case in point. The fact is that the term “industrial revolution” is so ingrained in our thoughts; it seems pointless to jettison it.

Despite such a historical achievement, a universal agreement on what constitutes an industrial revolution is still lacking [3]. However, from the perspective of the technological evolution [4], four fundamental general phases have been identified. In historical context, for example, the first industrial revolution in the early eighteenth century led to the discovery of the steam power and water which significantly increased the productivity of human labor by replacing manual handmade production. Furthermore, this innovation transformed society with trains, mechanization of manufacturing, a degree of automation, and of course smog. Almost 100 years later, the second industrial revolution evolved when electricity was notably a key driver. Mass industrial production led to productivity gains and opened the way for individualized mass consumption. Some 70 years later, the third industrial revolution led to the era of information technology (IT) resulting in the development and rise of computers and computer networks (WAN, LAN, MAN, etc.), the manufacturing of robotics, connectivity, and of course the birth of the Internet.

The fourth IR also referred to as industry 4.0 (4IR), which was coined by Klaus Schwab, started 30 years later and is arguably ongoing [5]. This specific industrial revolution combines technological and human capacities in an unprecedented way through self-learning algorithms, self-driving cars, human-machine interconnection, and big data analytics [6]. Differences between the fourth industrial revolution and the aforementioned three are clearly noticeable. In its scale, scope, and complexity, the transformation will be unlike anything humankind has experienced before. The fourth industrial revolution is not merely a prolongation of the third industrial revolution but rather a new and distinct revolution.

Brief History of AI

Why include the history of AI? This question is best answered by outlining its importance in any intellectual pursuit. History enables us to understand society of the past, thereby increasing our mastery over today's society [7]. This would lead to the views held by Tosh when he asserted that “To know about the past is to know that things have not always been as they are now, and by implication that they need not remain the same in future” [8, p2]. Thus far and based on the writings of these two authors, it would be relevant to conclude with what the renowned Spanish American philosopher, George Santayana, is famously quoted to assert that “Those who cannot remember the past are condemned to repeat it” [9].

Made not Born: The Beginning

Scholars can be forgiven for thinking that AI is the creation of modern minds. In fact, AI is not an innovation of current time. Its history can be traced back to Greek antiquity when such intellectuals from the time of Homer to Aristotle pondered over the idea of mechanical men and automatons. For example, nearly 3000 years ago in the *Iliad*, Homer described how Hephaestus created golden handmaidens and endowed them with reason and learning. Hephaestus also created

a bronzed automaton colossus tasked with defending the Minoan Kingdom of Crete against invaders and hurl boulders to sink any foreign vessels approaching Crete's shores [10]. It is said that Talos was killed by the attacked and killed by the Argonauts who removed the nail from around his ankle that sealed the outflow of blood. In so doing, the flow of his blood comprising of the "molten lava" was drained from his body leading to his death.

Greek mythology also cites the craftsmanship of Greek engineer Daedalus as being the first mortal to create "living statues" of bronze sculptures capable of demonstrating human characteristics whereby they could move their eyes, form tears, bleed, move, and speak. It is also said that Daedalus created a cow made of wood which on mating gave birth to the half-man and half-bull Minotaur. It is also suggested that Daedalus also made wings out of wax for himself and his son Icarus so they could escape from their prison in Crete. Icarus ignored warnings of flying close to the sun which melted his wings, thereby drowning in the sea.

Another noteworthy historical marking was the mythical Greek sculptor known as Pygmalion who created the beautiful statue of Galatea. Her beauty was such that Pygmalion fell so deeply in love with her that the gods brought Galatea to life. Later on, George Bernard Shaw reenacted in his theater play *Pygmalion* which became the famous musical *My Fair Lady*.

Moving forward to 1495, Leonardo Da Vinci is reputed to have designed an automaton robot with an anatomically correct jaw. The robot could stand, raise its visor, and independently move its arms. It is said that the operation of the entire system was based on a series of chains and pulleys. The discovery of his sketchbook in 1950 led to the successful replication of a fully functional robot.

In 1739, of his many creations, Jacques de Vaucanson of Grenoble, France, was well-known for the creation of the "robotic" duck known as the Vaucanson's Duck. This simulacrum of life was crucial for the enlightenment of automata construction. The use of perforated gold-plated copper plate used to build this life-sized duck

permitted a view of the functioning of the internal components. When activated, this creation moved like a duck, wiggled its beak in water, quacked, and readjusted its position. It was renowned for eating pellets and following a period of "digestion"; the pellets were excreted as fecal matter. Voltaire was most impressed and labeling him as the "new Prometheus" added: "sans le canard de Vaucanson, vous n'auriez rien qui fit ressouvenir de la gloire de la France" [11] (Without Vaucanson's Duck, you have nothing to remind you of the glory of France). Whether he was being sarcastic or not is left to the individual's thinking.

The intellectual legacies of early thinkers, philosophers, mathematicians, and logicians created the foundation for the development of "mechanized" humans. As a notable example, in 1928, Professor Makoto Nishimura, a Japanese biologist and botanist, built the first robot in Japan called *gakutensoku* which translates as "learning from the law of nature." This would imply that the robot could learn from people and nature.

The earlier thinkers were instrumental in ensuring that AI became increasingly more tangible throughout the 1700s and beyond. It became a time when philosophers contemplated how human thinking could be artificially mechanized and manipulated by intelligent nonhuman machines. The 1950s marked a momentous era when progress in AI underpinned by research findings truly came to fruition. The first major success in this decade was led by Alan Turing, a British mathematician among other things who proposed a test, the Turing test [12], that measured a machine's ability to replicate human actions to a degree that was indistinguishable from human behavior. The Turing test became a key component in the philosophy of AI as it discusses intelligence, consciousness, and ability in machines. However, the greatest success is arguably credited to John McCarthy, a British computer scientist who coined the phrase "artificial intelligence" in 1956 at the Dartmouth Conference. In 1958, McCarthy also developed a popular programming language called Lisp which became important in advancing machine learning (ML).

Following the conference and interest in AI, scientists were excited and began to make bold

statements and promises about the future findings of AI research. Consequently, sizeable research funding was received from governments and other organizations. Research and innovations in the field of AI began to flourish between 1957 and the 1960s when computers were faster and were able to store more information, thus demonstrating its impact on many real-world applications. However, a scientific discipline often encounters different challenges as it evolves and matures. With AI, it soon became evident that computational power was limited to solving small problems only. To progress it needed to overcome the challenges of AI research and to make the transition from being a “cool technology” to a “real-life presence” in human society [13]. Unfortunately, the promised outcomes failed to live up to the unreasonable hype, and the field of AI experienced intense critique. From the mid-1970s to the mid-1990s, the exuberance of AI soon gave way to a slump known as an “AI winter” when disillusionment set in and funding declined. After this phase of low point in the 1990s, optimism has since increased in AI and AI research funding.

Definition

The complexity involved in the development of synthetic intelligence as being comparable with human intelligence has led to varying interpretations of this technology. Some writers even seem reluctant to use the term AI preferring “computational intelligence” instead [14].

However, the possibilities of creating intelligent technologies have intrigued humankind since the inception of computers. Even the meaning of the word intelligence including AI has been central to much discussion and a source of confusion. Although a definition aims to influence the way forward, to date a universally agreed definition of “intelligence” or “intelligence” of AI is not in evidence [15]. However, Alan Turing, the founding father of AI, defined this technology as “the science and engineering of making intelligent machines, especially intelligent computer programs that can act as intelligently as human

beings” [12]. This would suggest that AI technologies resemble processes associated with human intelligence, such as reasoning, learning and adaptation, sensory understanding, and interaction—a concept popularized in such films as *The Terminator*; *I, Robot* (2004); *The Matrix*; *2001: A Space Odyssey* (1968); and *Ex Machina* (2015), which portrayed the fictive worlds profoundly altered by AI and automata. On the other hand, in the public discourse, AI is often characterized as sentient machines having human-like capabilities [16]. By inference, it describes a set of advanced technologies that enable machines to perform highly complex tasks effectively which would otherwise require intelligence that can’t be matched if a human being were to perform them [17].

Subsets of AI

Terminology surrounding AI technologies continues to evolve and can be a source of confusion particularly to a noncomputer scientist. In general terms, AI is broadly classified into machine learning (ML) and deep learning (DL) (Fig. 1). ML, a subset of AI, has been defined as the process of building computer systems that learn by analyzing huge amounts of data, apply algorithms to the data, and thereafter train themselves to make accurate informed decisions based on these underlying datasets rather than following pre-programmed rules [18]. As a rule, the operating process of a computer is underpinned by algorithms. This comprises a series of step-by-step instructions written by a programmer for software to follow verbatim. Algorithms, the building blocks for ML and AI, are almost everywhere around us. They comprise of a series of precise step-by-step instructions to do something or solve a problem. Following a cake recipe for making a cake or a nurse following a step-by-step procedure when undertaking a venepuncture is some of the many examples of algorithms. Another familiar example relates to nurses inputting keywords for a search strategy. The computer then processes the data in its database for items relevant to the words in the query. The results are the output which is

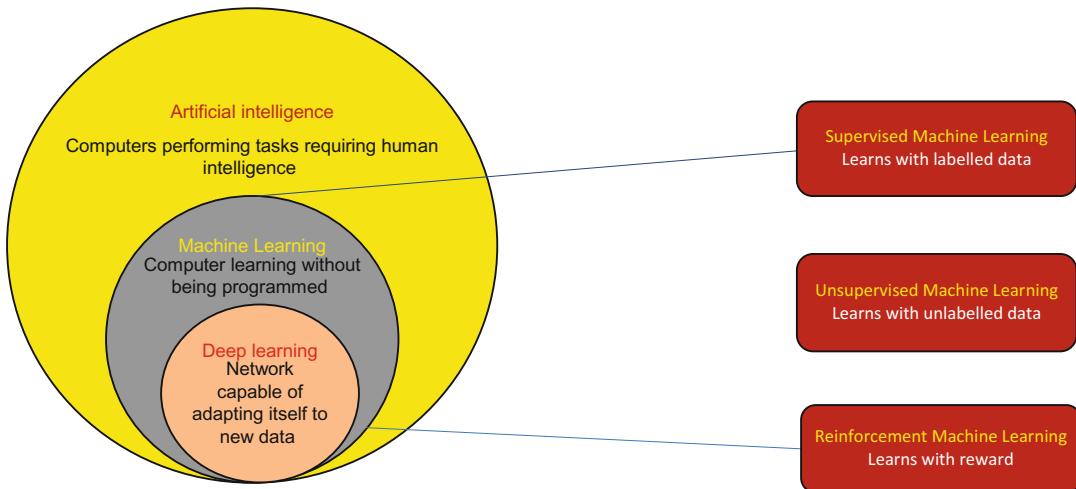


Fig. 1 Subsets of artificial intelligence

yielded through the ability of the computer to recognize patterns in the data.

Supervised Machine Learning (SML)

Based on Thomas Bayesian theorem and as the names suggest, SML (Fig. 2) involves machine learning algorithms (hereafter called the model), i.e., learning under the presence of a supervisor. The formal supervised learning process can be depicted by a simple formula, $Y = f(X)$, whereby the input variable is (X) and an output variable called (Y). The algorithm is used to learn the mapping function from the input to the output. In its basic mathematical context, the output (Y) is a dependent variable of input (X) represented by $Y = f(X)$. The ultimate aim is to approximate the mapping function (f) in order to predict the output variables (Y) during the input of new data. In reality if the mapping process is deemed correct, the algorithm has successfully learned; otherwise, necessary changes need to be made to the algorithm, thereby enabling it to learn correctly. SML algorithms are created to make predictions from future new unseen data.

The predicted labels can be both numbers and categories. For instance, if house prices are being predicted, then the yielded output will be a

number thus referred to as a regression model. If an email is being predicted as being spam or not, then the output is a category and falls under the rubric of a classification model.

A basic principle of SML can be exemplified with a scenario from the student nurse and teacher relationship in a university setting whereby the learning experience of the former is facilitated via the medium of textbooks, independent learning, and skill laboratory. If a test is set, then hopefully the student nurse will pass; otherwise, the marker's feedback will enable the student nurse to learn from his/her mistakes in order to successfully achieve the outcome.

Assuming you have a child who after graduating from learning the words mummy and daddy you decide to teach her what a dog or cat is. You could either show her/him videos of dogs and cats or related pictures or introduce her/him to a real-life dog and cat with the aim that the child also learns the differences and attributes between the two animals. If after a while the child can differentiate between the dog and the cat, then teaching has been successful. If the outcome was not achieved, then you will no doubt undertake some more coaching until a successful outcome is reached. In this example, you acted as a supervisor, and the child was the algorithm that needed to learn. This constitutes the basic principle of supervised learning.

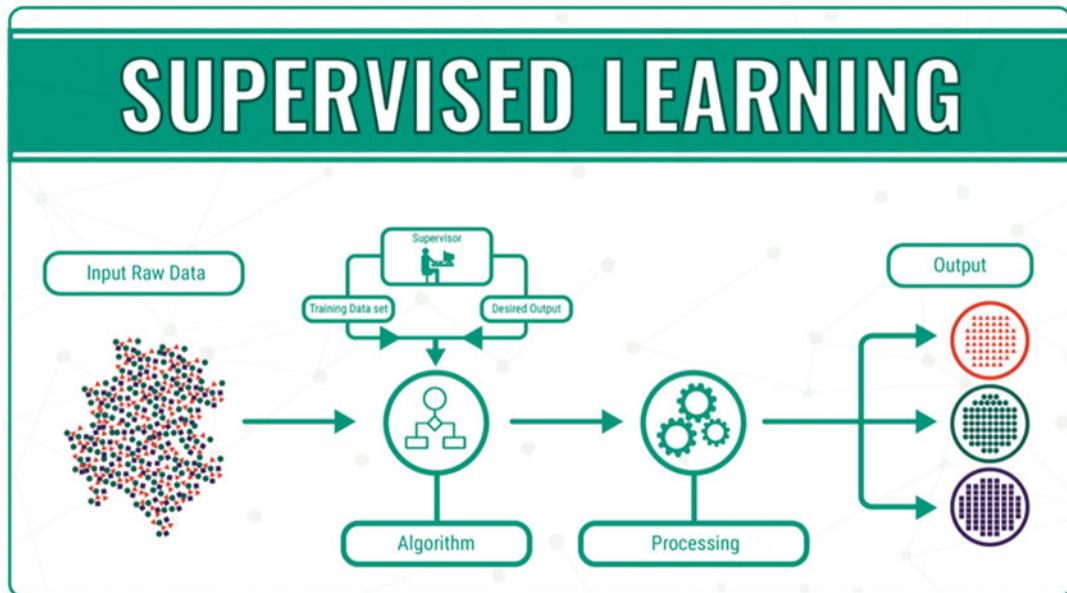


Fig. 2 Supervised machine learning (SML). (Source: With kind permission from Ronald van Noon)

Unsupervised Machine Learning (UML)

On seeing such movies as *The Terminator* or *Ex Machina*, it is probable that people have spoken about computers and machines possessing the ability to “teach themselves” in a seamless manner without the input from humans. In fact, this example suggests that in a way people are alluding to the processes involved in UML. Simply put, in UML (Fig. 3), a set of unlabeled data (raw data) is provided enabling a model to independently learn important structure relating to the dataset. Thereafter, the raw data is interpreted to identify and recognize hidden patterns in the dataset. This is followed by the application of a relevant algorithm which separates the data objects into groups based on similarities and differences between the objects and represents that dataset in a compressed format. The objective of UML is achieved through clustering when data are divided into distinct clusters based on distance to the centroid of a cluster or by association when the rules are used to discover interesting patterns.

Reinforcement Machine Learning (RML)

RML (Fig. 4), which is managed by a plethora of different algorithms, is a goal-oriented learning tool where a computer agent (algorithm), mapping and acting as an independent decision-maker, analyzes available data within its defined environment (task or simulation), derives a rule for taking actions (moves by agent), and maximizes long-term rewards or punishment (numerical values) through a trial-and-error system [19]. As a rule, the RML agent’s performance, which is not based on prior information, is rewarded or punished during each time step, thus allowing it to improve the performance of subsequent actions by trial and error [20].

A Case for ICU

It has been suggested [21, 22] that given the voluminous amount of data recoded from the complex comorbid nature of conditions experienced by

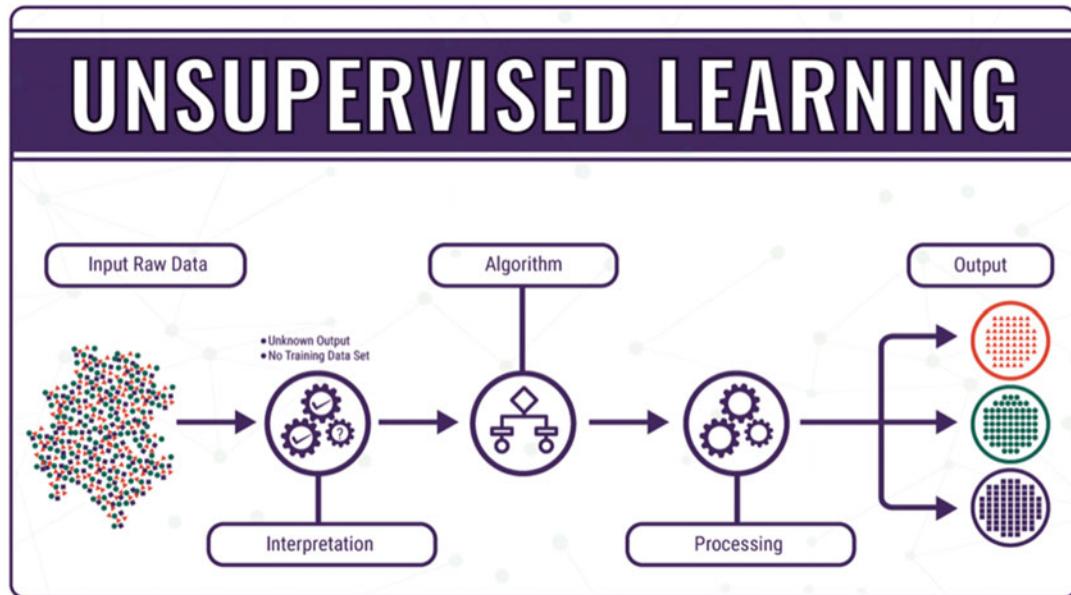


Fig. 3 Unsupervised machine learning (UML). (Source: With kind permission from Ronald van Noon)

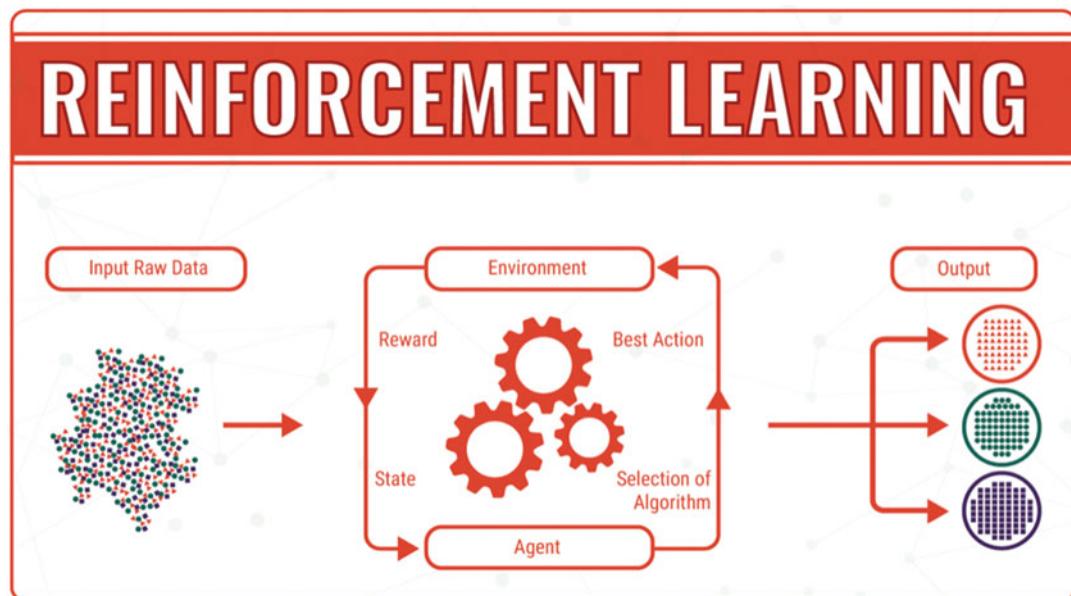


Fig. 4 Reinforcement machine learning (RML). (Source: With kind permission from Ronald van Noon)

patients admitted to ICU means that they are likely to benefit more from an RML approach system than by evidence-based clinical guidelines or best practices defined by physicians. Consequently, although nascent in nature, RML is still a well-

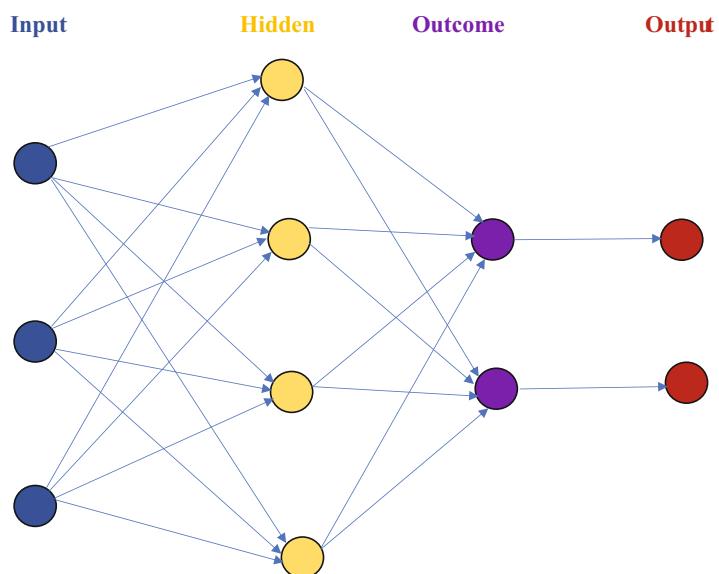
suited model for ICU since it can optimize treatment and patient outcome. Similarly, it can expand healthcare professionals' understanding of protocols and the potential to apply various treatment options.

The application of RML in ICU could be outlined with a simple example when evaluating a patient's state (demographics) and vital measurements (laboratory test results). After securing these data, a therapeutic intervention would be actioned followed by further tests to determine outcome. A reward would be assigned to the RML agent if the patient's state improves; otherwise, punishment would be assigned if the condition remains unchanged. In the latter instance, the physician would take another action reflective of the state. This state-action pairs for a patient which would roll out over time would represent the patient's condition, thus resulting as being fixed (first 24 h stay in ICU) or dynamic (when patient could be discharged at different times from ICU).

Deep Machine Learning (DML)

DML is, for example, a key technology behind driverless cars, enabling them to recognize a stop sign or to distinguish a pedestrian from a lamp-post. This technology is a subfield of ML with functional capabilities similar to ML. However, its functional properties are different. DL is concerned with algorithms inspired by the biological neural network and function of the human brain called artificial neural networks (ANNs). ANNs work by constructing layers upon layers

Fig. 5 Illustration of a simple neural network.
 (Source: Jiang et al. 2017)



of simple processing units (often referred to as “neurons”), interconnected via many differentially weighted connections [23] (Fig. 5). A key difference between the two technologies could be explained: thus, if ML algorithm returns an inaccurate prediction, then an engineer needs to step in to make necessary adjustments. But with a DL model, the algorithms can determine on their own if a prediction is accurate or not. As an example, ML could be programmed to recognize a hypoglycemic event based on the blood glucose parameters inputted. Eventually it could also pick up any phrase relating to a low blood glucose values reflecting the inputted parameters for hypoglycemia. However, a program with a DL model could recognize other variables, such as infection, stress, insufficient insulin or medication intake, and unscheduled strenuous physical exercise to mention a few, which can also contribute to hypoglycemia. This makes DL far more capable of carrying out complex tasks and making accurate predictions autonomously than machine learning.

Nursing

History informs us that the nursing profession evolved using the basic scientific principles of the nineteenth century [24], thus enabling nurses to have a critical role to be full-time hands-on care

providers. Throughout this time, nurses have also witnessed the influx of technological innovations. These have ranged from the “test-tube” babies, medical lasers, the artificial heart, genome mapping, CT and MRI imaging, angioplasty, dialysis, endoscopic procedures, bionic prosthetics, the Internet and health information technology (IT), the electronic health record (EHR), and robotic surgeries.

Despite nurses being the largest group of health professionals, the global healthcare system is experiencing a turbulent time. For example, staff shortage compounded by the continued decline in funding is unlikely to meet the ever-increasing demand driven by an aging population with multiple chronic conditions [25]. Bridging this gap between supply and demand will require changes rather than “simply throwing more resources at healthcare” [26]. Achieving this change means that a reform within the healthcare system is necessary if it is to continue delivering high-quality care underpinned by cutting-edge research. Reformation should also consider the implementation of a continued educational program for healthcare providers who face a huge knowledge challenge as the pace and complexity of medical knowledge now “exceeds the capacity of the human mind” [27]. Knowledge acquisition becomes a particularly important mechanism for nurses to strategically position themselves to remain relevant in the twenty-first century when healthcare is being ushered toward telemedicine and AI. This era of nursing-AI interface will potentially change the traditional approach to the nurse-patient relationship and the person and family-centered compassionate care. The success of this significant paradigm shift will invariably necessitate a strong and powerful nursing leadership to open discussion that will inform healthcare providers about opportunities and challenges for patient care and the profession.

Nursing Practice

Nursing is based on a framework known as the nursing process [28], comprising of a five-stage cyclical process known as assessment, nursing

diagnosis, planning, intervention, and evaluation [29]. Established in the early 1960s, this cutting-edge brainchild of Ida Orlando was depicted as a product that defined the nursing profession’s unique contribution to patient care. This interactive problem-solving approach became one of the signature roles of nurses in performing and achieving outcomes of nursing care. In practice, the nurse practitioner uses a degree of critical thinking to gather data, analyze and interpret the data, select, and apply appropriate intervention measures, concluding the process by evaluating the efficacy of the plan [30]. Thereafter, the development within the nursing profession led to nurses’ discovering and understanding the essence of evidence-based practice, defined as the “the conscientious, explicit and judicious use of current best evidence in making decisions about the care of the individual patient. It means integrating individual clinical expertise with the best available external clinical evidence from systematic research” [31]. Such an approach to care enabled the practitioner to apply a holistic approach to care that incorporated such elements as best research evidence, clinical expertise, the patient’s individual values, and circumstances, together with the characteristics of the practice in which the health professional works [32] and without excluding the use of experience, skills, training, patient’s values, and support [33]. The current process involved in the implementation of nursing interventions or actions to achieve a predictable outcome will arguably transcend within a technologically advanced future. Unless the present approach to care management which is underpinned by prescribed procedures is reviewed urgently, then the probability of such tasks being performed by humanoid nurse robots (HNRs) cannot be ignored. To avert this potentially likely scenario, it will be necessary that the future role of the nurse will be reviewed.

Societal Change

Arguably, advancements in the current exponential growth in robotics are being fuelled to also mitigate international shortage of nurses and to

support the current growing aging population in society brought about by the baby boomers [34]. It has been estimated that the current baby boomer generation in the United States, generally defined as the cohort born between 1946 and 1964 [35, 36], will reach 78 million by 2035, peaking to 100 million by 2060 compared to 76.4 million people under 18 [37]. As such this category of the elderly population will place greater demand on the healthcare system [38]. The consequences of an aging society, also known as Society 5.0, is more extreme in Japan as it is the only country in the world with the highest number of elderly people [39] and with a decrease in the birth rate, thereby negatively impacting on the figures entering the workforce [40]. The result of an aging Japanese society (*chokoreika*) combined with a reduced birth rate has led Japan to rely heavily on technology to provide human-centered care for the elderly people. It would therefore seem realistic and prudent to think that like Japan, the inclusion of humanoid nurse robots within the workforce could potentially resolve the strain being experienced in all sectors of the global healthcare system.

AI and Nursing Education

Historically the healthcare industry has always implemented machineries to complement care strategy. However, none of these technological innovations would be comparable with the capabilities and impact of the rapidly evolving AI in the care industry. Consequently, consideration should be given to the possibility that the integration of AI in healthcare is a likely cause of consternation among nurses. To address this state, it will be important to consider how the formal and informal educational programs and curricula could be re-examined to meet the needs of the existing and future nurses. The acquisition of new AI nursing competencies is most likely to be assured through the inclusion of such topics as basic health informatics knowledge and skills, concepts of data literacy, technological literacy, and AI algorithms together with a sound grasp of the ethical implications of using AI technology in the clinical context [41, 42]. The availability of

robots alongside the existing cadre of high-fidelity simulators in the school of nursing simulation laboratories is considered important in enabling nursing students to confidently gain hands-on experience with the technology [41].

The successful outcome of programs will also depend on the enhanced AI knowledge base of nurse educators teaching nurses in higher education [42]. Further, the desired outcome of any AI educational programs will also be underpinned by the involvement of a multidisciplinary approach to teaching that include information technologists, robotics engineers, and computer programmers. Such a collaborative approach which includes engineering principles will enhance nurses' understanding of technologies to be encountered in real-world clinical settings [34, 43]. The fact that nurses possess key insights into continually evaluating care to optimize outcomes makes them ideal candidates in co-designing technologies that address patients' needs and preferences programmed in nontechnical jargons that eliminates misunderstanding by patients [44]. The outcome likely to emerge from the co-creation of robots will enable the partnership between nurses and human nurse HNRs to deliver optimal patient care in a safer and a more effective and efficient manner.

Enhancing Holistic Care

This collaborative nurse-robot interface is more likely to strengthen nursing practice [45] rather than fearing that technology will replace nurses [46]. In welcoming innovation, nurses need to understand that both technology and the human nurse can each make a unique contribution in the delivery of competent of care. For example, AI technology can provide continuous real-time feedback about a patient's status than the moment-in-time assessment undertaken by the nurse. This combined data collation will provide a more accurate analysis of the patient's status, therefore a well-informed and approach to care. As nurses become more proactive in learning about AI technologies, HNRs in nursing will be valued as key members of the multidisciplinary care team. In turn, automation of repetitive tasks

will enable human nurses to focus more on the art of caring by building relationships, exercising empathy, and using human judgment to guide and advise.

Robots to the Rescue

Today society is witnessing an unprecedented rate of evolutionary and incremental as well as revolutionary and transformative technological development within nursing and the healthcare industry. For example, it is becoming evident that AI-driven robots are increasingly prevalent in the clinical settings worldwide. These robots (Table 1) have improved tasks and the procedural functionalities that are being performed in a more efficient and safer manner.

Table 1 Examples of technological innovations with functional properties

Robotic technology	Function
Paro	Nuzzles those who stroke or talk to it
Robear and RIBA	Capable of transferring patients from a bed to a chair
da Vinci surgical robot	Used in various types of surgeries where surgeons control the system and directs procedure
Cody	Gives bed baths to patients
Veebot	Has an 83% accuracy in selecting the best vein
Robotic prescription dispensing systems	Made the dispensing of pharmaceuticals more accurate and safer. Decreases the responsibilities of nurses in medication administration
Lynx Autonomous Intelligent vehicles	Capable of moving goods in a large facility and can self-navigate in dynamic environments
Swisslog RoboCourier	Used for transport of specimens, medications, and supplies in hospitals, clinical laboratories, and pharmacies
Xenex	Uses high-intensity ultraviolet light to disinfect any surfaces of the healthcare facility
Aibo	A friendly pet for the elderly
Pepper	Specializes in customer service, monitors corridors at night, and communicates with patients
Chapit	Engages in elementary conversations
Palro	Able to lead a group of elderly people in an exercise routine
Tug	Autonomous mobile delivery system that transports equipment weighing up to 450 kg
Ribo and Buddy	Other than being a friend, recognize individual family members of a household, take photos, read books to children, assist in the kitchen, relay messages, and share weather forecast
SARA	A Dutch design robot to support nursing staff in taking care of the elderly. SARA can also, for example, assist the elderly with their exercises, tell them stories, or warn the nurses if something goes wrong
Baxter	A British-built robot that can detect if a human needs help moving or dressing and use its sensors and dexterous fingers to lend a hand
ElliQ	California, a robotic companion for the elderly
Dinsow	Assists with mood, improving activity, and acts as a reminder to take medication

The Future of Robotics

Just like the rise of robotics in nursing, technology is developing at a fast rate within the medical arena. In a continued effort to improve healthcare, the future of surgery is gradually experiencing a radical change. For example, some of the robotics which are in their early stage of development hold the promise to assist in surgery (Table 2), thus improving direct patient care.

Conclusion

A noticeable decline in funding growth in a challenging global economy along with an international shortage of healthcare workforce is unlikely to meet the increasing demand driven

Table 2 Robotics of the future

Robotic technology	Function
Medrobotic's Flex Robotic System	This robotic technology has shown some success in assisting in the visualization in general surgical, gynecological, and thoracic procedures in the United States
Hansen Flex Robotic System	This technology is used during electrophysiology procedures by remotely navigating a robotic catheter in the cardiac atria and can also provide 3D catheter control including 3D visualization during these procedures
Medtech – ROSA Spine	Can perform minimally invasive spinal surgery
Atheon – TUG	This autonomous mobile delivery robot is capable of transporting racks, carts, or bins weighing up to 543 kg

particularly by an aging baby boomer generation living with multiple complex chronic health conditions. To resolve the problem, attention has focused on AI technology to contribute toward delivering healthcare as well as to boost productivity and bring about significant economic and social benefits.

Today, our world has entered a period of truly transformative change propelled by the pace and scale of technological developments that was simply not anticipated. Such an avalanche of technological advances will undoubtedly reshape the very essence of humanity and touch every aspect of life on the planet. In the context of healthcare, nurses should not simply focus on how technology detracts from humanistic care. Instead, it will be more appropriate for them to explore how healthcare technology and compassionate care play key roles in the provision of compassionate nursing practice. Further, nurses need to reconsider their roles by analyzing how the inclusion of AI in care can enhance critical thinking and the nursing process. This collective approach to delivering care underpinned by evidence will lead them to support the patient's journey adding value through expedited, more precise, and enriched decision-making. As nurses' knowledge of AI increases, they will be able to delegate technical roles and repetitive tasks to HNRs, thereby enabling them to spend more time with patients and families. To summarize, it is only through gaining AI knowledge that nurses will realize the potential of this technology within the healthcare settings. So rather than alienating themselves from this technology, nurses should infuse AI to optimize healthcare outcome in a safer and more efficient manner. Irrespective of whether AI

singularity will come to pass or not, AI in healthcare sectors will continue to improve by overcoming challenges, and these improvements which are continually being achieved appear to be accelerating.

The notion that AI might create existential risk to humanity has certainly fuelled currency by prominent thinkers [47] and public intellectuals [48] as well as the creation of a dystopian society as portrayed by Hollywood movies. Although the age of AI has arrived, the question of whether people should fear this technology remains. The AI-induced apocalyptic scenarios presented to us in many films and television series depicting AI gone wrong have certainly cultivated a sense of fear and apprehension. Moving away from the cinematic space, AI is revealing endless possibilities of serving humanity in all its enterprises [49]. Finally, whether the depiction of a dystopian society is motivating, frightening, or discouraging is difficult to assert, but at least they have posited the ethical dimension of technology on the agenda. On a final note, while celebrating the present and future success of AI, we should take heed of the cautionary wisdom of Eliezer Yudkowsk [50] (2008) when he said that "By far the greatest danger of Artificial Intelligence is that people conclude too early that they understand it."

References

1. Cambridge Dictionary. The industrial revolution. 2017. <https://dictionary.cambridge.org/dictionary/english/industrial-revolution>. Retrieved 22 Feb 2021.
2. Belvedere V, Grando A, Bielli P. A quantitative investigation of the role of information and communication

- technologies in the implementation of a product-service system. *Int J Prod Res.* 2013;51(2):410–26.
3. Maynard AD. Navigating the fourth industrial revolution. *Nat Nanotechnol.* 2015;10(12):1005–6.
 4. National Academy of Science and Engineering – ACATECH. Recommendations for implementing the strategic initiative industrie 4.0. Final report of the industrie 4.0 working group. Frankfurt: ACATECH Report; 2013.
 5. Miller D. Natural language: the user interface for the fourth industrial revolution. Opus Research Report. 2016.
 6. Schaffer M. The fourth industrial revolution: how the EU can lead it. *European View.* 2018;17(1):5–12. <https://doi.org/10.1177/1781685818762890>.
 7. Carr EH. What is history? London: Penguin Books; 1987.
 8. Tosh J. The pursuit of history. London: Longman; 1984.
 9. Santayana G. The life of reasons. New York: Prometheus Books; 1905.
 10. Mayor A. Gods and robots. Princeton University Press: Princeton; 2018.
 11. Voltaire (pseud van François-Marie Arouet). *Oeuvres complètes de Voltaire.* (in French). P. Paris: Plancher; 1819. p. 491.
 12. Turing AM. Computing machinery and intelligence. *Mind.* 1950;59:433–60.
 13. Lighthill J. Artificial intelligence: a general survey. In: Artificial intelligence: a paper symposium. 1973. p. 1–21.
 14. Poole DL, Mackworth A, Goebel RG. Computational intelligence and knowledge. In: Computational intelligence: a logical approach. New York: Oxford University Press; 1998. p. 1–22.
 15. Monett D, Lewis CWP. Getting clarity by defining artificial intelligence – a survey. In: Muller VC, editor. *Philosophy and theory of artificial intelligence* (2017). Berlin: Springer; 2018. p. 212–4.
 16. The Royal Society. Machine learning: the power and promise of computers that learn by example. London: The Royal Society. 2017. <https://royalsociety.org/~media/policy/projects/machine-learning/publications/machine-learning-report.pdf>. Accessed 25 Feb 2021.
 17. Hall W, Pesenti J. Growing the artificial intelligence industry in the UK. Report. London: HM Government; 2017.
 18. Kassahun Y, Yu B, Tibebu AT, et al. Surgical robotics beyond enhanced dexterity instrumentation: a survey of machine learning techniques and their role in intelligent and autonomous surgical action. *Int J Comput Assist Radiol Surg.* 2016;11:553–68.
 19. Sutton RS, Barto AG. Reinforcement learning: an introduction. 2nd ed. The MIT Press: Cambridge, MA; 2018.
 20. Kiumarsi B, Vamvoudakis KG, Modares H, Lewis FL. Optimal and autonomous control using reinforcement learning: a survey. *IEEE Trans Neural Netw Learn Syst.* 2018;29(6):2042–62. <https://doi.org/10.1109/TNNLS.2017.2773458>.
 21. Chen Z, Marple K, Salazar E, Gupta G, Tamil L. A physician advisory system for chronic heart failure management based on knowledge patterns. *Theory Pract Logic Program.* 2016;16(5–6):604–18. <https://doi.org/10.1017/S1471068416000429>.
 22. Almirall D, Compton SN, Gunlicks-Stoessel M, Duan N, Murphy SA. Designing a pilot sequential multiple assignment randomized trial for developing an adaptive treatment strategy. *Stat Med.* 2012;31(17):1887–902. <https://doi.org/10.1002/sim>.
 23. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol.* 2017;2:e000101. <https://doi.org/10.1136/svn-2017-000101>. 230–243.
 24. Buhler-Wilkerson K, D'Antonio P. History of nursing. *Encyclopedia Britannica.* 2019. <https://www.britannica.com/science/nursing>. Accessed 28 Feb 2021.
 25. Department of Health. Policy paper: 2010 to 2015 government policy: long term health conditions. 2015. <https://www.gov.uk/government/publications/2010-to-2015-government-policy-long-term-health-conditions/2010-to-2015-government-policy-long-term-health-conditions>. Accessed 28 Feb 2021.
 26. Milburn A. Technology and innovation are key to saving the NHS. *The Guardian.* 22 February 2021. 2017.
 27. Obermeyer Z, Lee TH. Lost in thought – the limits of the human mind and the future of medicine. *N Engl J Med.* 2017;377(13):1209–11.
 28. Orlando I. Nursing in the 21st century: alternate paths. *J Adv Nurs.* 1987;12(4):405–12.
 29. Alfaro-Lefevre R. Nursing process overview. In: Kogut H, editor. *Applying nursing process.* 6th ed. Philadelphia: Lippincott Williams & Wilkins; 2006. p. 4–41.
 30. Muller-Staub M, Lavin M, Needham I, Archterberg T. Nursing diagnoses, interventions and outcomes – application and impact on nursing practice: a systematic review. *J Adv Nurs.* 2006;56(5):514–31.
 31. Sackett D, Rosenberg W, Gray J, et al. Evidence based medicine: what it is and what it isn't: it's about integrating individual clinical expertise and the best external evidence. *Br Med J.* 1996;312:71–2. <https://doi.org/10.1136/bmj.312.7023.71>.
 32. Mayer D. *Essential evidence-based medicine.* 2nd ed. Cambridge University Press: Cambridge; 2010.
 33. Hoffman T, Bennett, Del Mar C (2013). *Evidence-based practice: across the health professions* (2nd ed.). Elsevier: Chatswood.
 34. Maalouf N, Sidaoui A, Elhajj IH, Asmar D. Robotics in nursing: a scoping review. *J Nurs Scholarsh.* 2018;50(6):590–600.
 35. Haaga J. Just how many Baby Boomers are there? 2002. <http://www.prb.org/Articles/2002/JustHowManyBabyBoomersAreThere.aspx>. Accessed 25 Feb 2021.

36. Rogerson PA, Kim D. Distribution and redistribution of the baby-boom cohort in the United States: recent trends and implications. *Proc Natl Acad Sci U S A.* 2005;102(43):15319–24. <https://doi.org/10.1073/pnas.0507318102>.
37. Song Z, Ferris TG. Baby boomers and beds: a demography challenge for the ages. *J Gen Intern Med.* 2018;33(3):367–9.
38. McCarthy J, Minsky ML, Rochester N, Shannon CE. A proposal for the Dartmouth summer research project on artificial intelligence. *Artif Intell Mag.* 2006;27(4):12–4.
39. Chen BK, Jalal H, Hashimoto H, Suen SC, Eggleston K, Hurley M, Schoemaker L, Bhattacharya J. Forecasting trends in disability in a super-aging society: adapting the future elderly model to Japan. *J Econ Ageing.* 2016;8:42–51.
40. Elsy P. Elderly care in the society 5.0 and Kaigo Rishoku in Japanese hyper-ageing society. *Jurnal Studi Komunikasi.* 2020;4(2):435–52. <https://doi.org/10.25139/jsk.v4i2.2448>.
41. Health Education England. The Topol review: preparing the health care workforce to deliver the digital future [White paper]. 2019. <https://topol.hee.nhs.uk/wpcontent/uploads/HEE-Topol-Review-2019.pdf>. Accessed 22 Feb 2021.
42. Tanioka T, Yasuhara Y, Dino MJS, Kai Y, Locsin RC, Schoenhofer SO. Disruptive engagements with technologies, robotics, and caring: advancing the transactive relationship theory of nursing. *Nurs Adm Q.* 2019;43(4):313–21.
43. Glasgow MES, Colbert A, Viator J, Cavanagh S. The nurse-engineer: a new role to improve nurse technology interface and patient care device innovations. *J Nurs Scholarsh.* 2018;50(6):601–11.
44. Backonja U, Hall AK, Painter I, Kneale L, Lazar A, Cakmak M, Thompson HJ, Demiris G. Comfort and attitudes towards robots among young, middle-aged, and older adults: a cross-sectional study. *J Nurs Scholarsh.* 2018;50(6):623–33.
45. Archibald MM, Barnard A. Futurism in nursing: technology, robotics and the fundamentals of care. *J Clin Nurs.* 2018;27(11–12):2473–80.
46. Locsin RC, Ito H. Can humanoid nurse robots replace human nurses? *J Nurs.* 2018;5(1):1–6.
47. Bostrom N. Superintelligence: paths, dangers, strategies. Oxford: Oxford University Press; 2014.
48. Hawking S, Russell S, Tegmark M, Wilczek F. Transcendence looks at the implications of artificial intelligence – but are we taking AI seriously enough? *The Independent*, 1 May 2014.
49. Smith S. The future of artificial intelligence. In: eLearning systems. eLearning Industry; 2016.
50. Yudkowsky E. Artificial intelligence as a positive and negative factor in global risk. In: Bostrom N, Ćirković MM, editors. Global catastrophic risks. New York: Oxford University Press; 2008. p. 308–45.



Nilakash Das, Marko Topalovic, and Wim Janssens

Contents

Introduction	760
Imaging	760
Chest X-ray	761
Computed Tomography (CT)	762
Histopathology	763
Bronchoscopy	763
Point-of-Care Ultrasound (PoCUS)	763
Lung Function Testing	763
Pulmonary Function Tests	764
Spirometry	764
Forced Oscillation Tests	764
Telemedicine	764
Miscellaneous	765
Sleep Monitoring	765
Breath Analysis	765
Lung Sound Analysis	766
Conclusions and Future Perspectives	766
References	767

Abstract

N. Das (✉) · W. Janssens

Laboratory of Respiratory Diseases and Thoracic Surgery,
Department of Chronic Diseases, Metabolism and Ageing,
Katholieke Universiteit Leuven, Leuven, Belgium

e-mail: neel.das@kuleuven.be;
wim.janssens@kuleuven.be

M. Topalovic

Laboratory of Respiratory Diseases and Thoracic Surgery,
Department of Chronic Diseases, Metabolism and Ageing,
Katholieke Universiteit Leuven, Leuven, Belgium

ArtiQ NV, Leuven, Belgium
e-mail: marko.topalovic@artiq.eu

Historically, the field of respiratory medicine has been a breeding ground for pioneering applications of artificial intelligence (AI). Recently, the field has seen an explosion of interest in AI applications that has been primarily driven by advances in computing power, algorithmic innovations, and availability of large datasets. In this chapter, we examine the applications of AI across different modalities of respiratory medicine, such as chest imaging,

lung function testing, telemedicine, sleep medicine, etc. We provide a historical context of these applications as well as highlight the latest trends, including the remarkable successes of deep neural networks (DNNs). Finally, we share our future perspective while pointing out the existing barriers to the implementation of AI systems in routine clinical practice.

Keywords

Respiratory Medicine · Artificial intelligence · Healthcare · Medical imaging · Computed tomography · Chest X-rays · Pulmonary function testing · Spirometry · Telemonitoring · Medicine

Introduction

Technological advances have always pushed the boundaries of what is possible in respiratory medicine. Since the development of spirometry and chest radiography in the nineteenth century, the diagnosis, treatment, and management of respiratory diseases have radically improved over the last century. Today, we are witnessing the genesis of another technological revolution brought about by artificial intelligence (AI).

Specifically, we define AI as computer algorithms simulating intelligence for tasks that seem intuitive to humans, such as reasoning, knowledge representation, learning, natural language processing, perception, and the ability to manipulate physical surroundings [1]. Machine learning (ML) is a subset of AI in which the computer automatically learns a task from predefined data features. Within ML, deep learning (DL) is a class of algorithms where the computer extracts the necessary features from raw input data. Deep learning models or deep neural networks (DNNs), such as convolution neural network (CNN) and recurrent neural network (RNN), have become very popular lately due to the availability of large labeled datasets, efficient training algorithms, and advances in parallel processing [2].

The field of respiratory medicine has been a breeding ground for many pioneering applications of AI in healthcare. It has been part of the

formative years of AI as expert systems to interpret pulmonary function tests (PFTs) and in its transformation to the first machine learning and deep learning applications on medical imaging data [3, 4]. In the last several years, there has been an explosion in interest in medical AI applications, a trend also observed in the field of AI in respiratory medicine. The number of articles published in just the last year alone constituted 90% of all articles published since 2000 [5].

In this chapter, we review the advances in ML and DL as applied to respiratory medicine. We will describe the history and illustrate the latest trends across different clinical modalities such as imaging, lung function testing, critical care monitoring, sleep monitoring, and telemedicine. Finally, we discuss the future perspective of AI in relation to respiratory medicine.

Imaging

Almost half a century back in 1966, Gwilym S. Lodwick introduced the term “computer-aided diagnosis” in scientific literature. He emphasized, “there is scarcely any repetitive function in which the computer cannot be of help to us, in radiology” [6]. Three years earlier, he had developed a visually descriptive system to convert X-ray images into numerical sequences, which in today’s ML terminology are called feature vectors. He postulated that these sequences could be manipulated by a digital computer to predict a diagnosis [7] and further demonstrated a relationship between these features and one-year survival of advanced lung cancer patients.

Scanning and processing radiographic images in the computer memory finally became possible in the 1970s. Image processing typically involved low-level operations such as filtering for detecting edges and lines, fitting simple mathematical structures, etc. [4]. At that time, rule-based methods that employed step-by-step procedures on low-level imaging features for finding lungs, heart, ribs, and, finally, abnormal regions were in fashion [8, 9]. However, these methods were often too brittle, similar to expert systems.

To address these challenges, the trend shifted to textural analysis, which marked the entry of the modern ML paradigm in chest imaging [10]. In textural analysis, a discrete region of interest (ROI) within the lung field is selected and a feature extraction algorithm based on structural [11], statistical, or binary patterns is applied [12]. Then a supervised learning algorithm is used to classify the textures. The group of Kunio Doi at the University of Chicago published many seminal studies in this field [13], and until recently, this was the dominant approach. A major weakness of such approaches is that it is impossible to determine which features are optimal. This makes them strongly application-dependent as it requires prior knowledge of the characteristics of the texture [4].

This is where deep learning entered the scene by completely taking over the task of both learning the features and classifying the problem. By removing the human element altogether, it marks a paradigm shift in the history of computational medical imaging. The trend started precisely in 2012 when Krizhevsky et al. trained a CNN with an error rate that was less than 10.8 percentage points than the runner-up in the ImageNet challenge, an annual computer vision competition organized by Stanford University [14]. Until then, training DNNs were computationally expensive, but it became feasible due to the development of a parallel processing framework with graphical processing units (GPUs). Today, the application of DNNs to chest radiographs and CT scans has resulted in a step change in diagnostic accuracy when compared to ML algorithms on semantic features such as tumor speculation or quantitative features such as shape and texture [5].

In this section, we will review the latest applications of DNNs within different medical imaging modalities.

Chest X-ray

Chest X-ray is the most commonly requested radiologic examination. Analysis of chest X-rays is formulated as a computer vision

problem in which CNNs have become state of the art. CNNs have been trained to detect abnormal findings on plain chest X-rays and specific pathologies, including tuberculosis [15, 16], pneumonia [17], and malignant pulmonary nodules [18]. Rajpurkar et al. developed a CNN to detect the presence of 14 different pathologies, including pneumonia and pulmonary masses and nodules [19]. While no significant differences in the area under the receiver operating characteristics curve (AUROC) was observed between radiologists and the AI, the time to interpret the images was significantly longer for the radiologists (240 min vs 1.5 min).

In a similar study, Hwang et al. developed a parallel network of CNNs to detect abnormalities associated with major diseases like lung cancer, tuberculosis, pneumonia, and pneumothorax on chest radiographs, as well as to localize the abnormalities. While the AI consistently outperformed thoracic radiologists and physicians in identifying disease abnormalities ($AUROC = 0.983$ vs $0.814\text{--}0.932$, $p < 0.05$) [20], the humans significantly improved their diagnostic performance when they assessed the radiographs with the assistance of the AI. Further, the same group demonstrated that the same AI improved the sensitivity of radiology residents in the detection of clinically relevant abnormalities in emergency room settings [21].

A common application of AI is to triage chest X-rays in clinical workflows. Annarumma et al. trained a CNN to triage chest radiographs as “normal,” “non-urgent,” “urgent,” and “critical” using a dataset of 329,698 images [22]. With a sensitivity and specificity of 71% and 95%, respectively, the AI further resulted in a fourfold reduction in the time to report radiographs with critical findings and a twofold reduction in the time to report urgent findings. During the COVID-19 pandemic, triaging of patients using AI on chest X-rays to detecting pneumonia-related abnormalities received a lot of attention [23, 24]. As this is a rapidly evolving situation, these results should be considered with a lot of caution due to uncertainties in data reporting and the high risk of bias [25].

Computed Tomography (CT)

CT is an advanced imaging technique that provides an insight into the lung structure in vivo. It is an ideal tool for the assessment of airway diseases and for screening lung cancer. DNNs have been extensively used in CT-based diagnostic workflows in respiratory medicine. Gonzalez et al. trained a CNN on 7983 CT scans from COPDGene, a large well-characterized cohort of airflow obstructed subjects, to stage the severity of chronic obstructive pulmonary disease (COPD) [26]. The same model was also used in predicting acute respiratory disease (ARD) events and mortality in smokers. Using the same cohort, Humphries et al. developed a CNN to classify the presence and severity of emphysema, which concurred well with visual scores of emphysema [27]. The CNN scores also contributed to the prediction of mortality and lung function parameters in COPD.

The availability of public CT datasets has precipitated a growing interest in applying DNNs to identify patterns associated with interstitial lung disease (ILD) such as fibrosis, consolidation, ground glass opacity, etc. [28]. In 2018, Walsh et al. published a seminal study in the application of DNNs for diagnosing ILD [29]. The group trained their model on 420,096 unique four-slice montages created from 929 CT scans. In the internal test set with 139 scans, the model demonstrated an accuracy of 76.4%. In another test dataset ($N = 150$ CT scans), the model outperformed 60 out of 91 thoracic radiologists in the diagnosis of fibrotic lung disease with a diagnostic accuracy of 73.3%. A recent study by Bermejo-Pelaez et al. developed a DNN to detect subtle interstitial changes that precede the development of ILD [30].

DNNs have also been used in the detection of pulmonary embolism. Tajbakhsh et al. carried out a comparative investigation to detect pulmonary embolism using a pretrained CNN with fine-tuning (transfer learning), which outperformed a CNN built from scratch, even at its worst case [31]. DNNs have also been applied to detect and calculate clot burden of acute pulmonary embolism on CT pulmonary angiography, which further correlated with measures of ventricular function [32].

There has been accumulating evidence that the implementation of low-dose CT can reduce mortality in lung cancer. An important limiting factor in this implementation is the availability of radiologists to report the large volume of screening CT scans. There has therefore been substantial interest in developing AI systems that can detect and accurately diagnose malignant pulmonary nodules on CT imaging. While the traditional approach involved applying ML on radiomics features from CT [33, 34], the trend has mostly shifted toward the application of DNNs for end-to-end screening [35–37]. Ardilla et al. developed a DNN using cases from the National Lung Cancer Screening Trial, which outperformed six board-certified radiologists when the current CT scan was available and was equivalent to the radiologists when both current and past scans were available for review [35]. Shen et al. developed an interpretable model that provided low-level semantic features often reported by radiologists with representations learned by the model, to explain predictions of nodule malignancy [37].

Researchers have also shown that DNNs can be used to predict prognosis and tumor type based on CT images. Hosny et al. trained a DNN to predict survival based on CT appearances in patients with non-small cell lung cancer undergoing surgery or radiotherapy [38]. The DNN distinguished between early (<2 years) and late (≥ 2 years) mortality in patients undergoing surgery and radiotherapy. Wang et al. reported that a DNN could predict epithelial growth factor receptor mutation status in patients with lung adenocarcinoma based on CT images [39]. The accuracy of the DNN significantly exceeded that of predictive models using clinical, semantic, or radiomics features.

During the COVID-19 pandemic, several researchers used DNNs to differentiate COVID-19 pneumonia from common pneumonia on CT scans [40, 41], and to quantify lung abnormalities [42]. The premise was that COVID-19 pneumonia lungs present with pronounced ground glass opacities that can be identified by a DNN [43]. These methods were mostly developed to identify high-risk patients with poor prognosis for early

intervention or for optimizing medical resources in emergency room settings.

Histopathology

Visual inspection of histopathology slides is one of the main methods used by pathologists to assess the stage, type, and subtype of lung tumors. Many AI-based diagnostic systems that provide an automated assessment of tissue images have been explored to aid diagnostic workflows in histopathology. Couدرay et al. trained a CNN on whole slide images from the Cancer Genome Atlas to automatically classify tumor types into adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC), which showed a comparable performance to expert pathologists [44]. Further, the model also predicted ten most commonly mutated genes in LUAD. Similarly, Shia et al. developed a DNN to predict tumor programmed ligand 1 (PD-L1) status from whole slide images of non-small cell lung cancer [45]. DNNs have also been applied to differentiate between lung adenocarcinoma growth patterns (acinar, micro-papillary, solid, lepidic, and cribriform) [46, 47], as well as to detect lung cancer metastases in lymph node slides [48]. Courtiol et al. developed a CNN on whole slide images to predict the overall survival of mesothelioma patients, without any pathologist-provided locally annotated regions [49]. They explained the predictions by highlighting regions that were mainly located in the stroma and are histological features associated with inflammation, cellular diversity, and vacuolization.

Bronchoscopy

Bronchoscopic inspection, as a follow-up procedure after radiological imaging, plays an important role in the diagnosis and treatment of lung disease patients. Bronchoscopy is followed by a biopsy, which requires the pulmonologist to be very selective as it may cause life-threatening haemorrhage. In this regard, DNN techniques on bronchoscopic images could provide objective recommendations on lung pathology and enable clinicians to be

more selective. However, the lack of large datasets prohibits the training of DNNs from scratch. Tan et al. fine-tuned an existing DNN (transfer learning) on a small sample of lung bronchoscopic images to differentiate between cancer, tuberculosis, and normal cases with an overall accuracy of 82% [50]. Feng et al. extracted textural features from bronchoscopic images and fed them into an ML classifier to distinguish between LUAD and LUSC [51].

Point-of-Care Ultrasound (PoCUS)

Emergency physicians often have access to PoCUS, a portable diagnostic tool that helps in minimizing the delay between the onset of symptoms and the initiation of definitive therapy [52]. Although literature on the application of AI on PoCUS imaging remains limited, some researchers have applied DNNs to detect the presence of B-line artifacts and further classify its severity [53]. More recently, Roy et al. used a DNN that scored and localized the severity of COVID-19 on PoCUS images [54].

Lung Function Testing

Pulmonary function tests (PFTs) are considered an important tool for the evaluation of the respiratory system and for diagnosing obstructive diseases such as asthma, chronic obstructive pulmonary disease (COPD), and lung fibrosis. Traditionally, we refer to complete PFT as the combination of spirometry, whole-body plethysmography, and diffusion capacity test, although tests like forced oscillation test (FOT) and nitrogen washout also fall under the purview of lung function testing. The interpretation of lung function testing has been considered as an important aspect among respiratory physicians.

There are several guidelines to aid in the differential diagnosis of common respiratory diseases [55, 56]. Although guidelines seem to be helpful in defining the typical patterns based on strict cutoffs, many cases have mild and varying disturbances that do not fit in any of the predefined boxes. As a result, the interpretation of PFT

requires expert knowledge, which is expensive and not always available at all levels of healthcare. Moreover, additional tests are often requested to confirm a diagnosis.

The potential of AI was recognized as a tool to provide an automated interpretation of PFTs by reproducing the cognitive abilities of a physician. Early attempts were made in the 1980s using rule-based expert systems such as PUFF and Pulmonary Consult [3, 57], which could model the interpretation pattern of an individual. However, such systems were brittle, and coupled with the limitations on computing power at the time, they never gained any widespread adoption.

After a long hiatus, there has been a resurgence of AI for PFT interpretation in the recent years. This has been largely motivated by the recent successes of machine learning in various applications across the medical domain. It was realized that machine learning can assist in providing a faster and more accurate diagnostic interpretation of PFT because of its ability to find patterns in a high-dimensional feature space [28].

In this section, we summarize the applications of AI and ML across different lung function testing methods.

Pulmonary Function Tests

Topalovic et al. developed an ML model using 1430 PFT cases that could differentiate between eight types of respiratory disease. In a head-to-head comparison using 50 test cases, the model demonstrated an accuracy of 82% and outperformed 120 European pulmonologists by a wide margin [58]. Machine learning has also been applied to detect ILD in systemic sclerosis patients using PFT [59].

Spirometry

Spirometric forced expiration lies at the foundation of PFT testing, and several researchers have applied ML techniques on spirometry data to extract clinical biomarkers. Bodduluri et al. developed a DNN to identify quantitative CT-based structural phenotypes

on the COPDGene cohort [60]. The model was more accurate in differentiating emphysema/small-airway phenotypes than traditional spirometric measures like forced expiratory volume in 1 s (FEV₁) and Tiffeneau index. Kaplan et al. explored different ML models and clinical features to differentiate asthma, COPD, and very subtle asthma-COPD overlap syndrome [61]. The group utilized a very large cohort of around 400,000 subjects extracted from US electronic health records covering primary care, specialist care, and hospital medical records. The best model featured spirometry measurements with BMI, pack years, symptoms, and history of allergic and chronic rhinitis as input features. Recently, Das et al. trained a CNN to automate the process of visual evaluation spirometric flow-volume loops, which is a source of large intertechnician variability in spirometric quality control [62].

Forced Oscillation Tests

The forced oscillation test (FOT) is a noninvasive and effort-independent method to measure respiratory impedance using sound waves. It offers a more accurate measurement of central and peripheral airway obstruction than spirometry [63]. FOT generates many frequency-dependent measurements, and an ML approach may help in uncovering different phenotypes in such a rich feature set. The group of Amaral et al. have been extensively involved in applying ML techniques in detecting COPD and categorizing severity in COPD patients [64, 65]. They also developed an ML approach to assist in the early diagnosis of smoking-induced respiratory changes [66]. Recently, they compared several classifiers to detect airway obstruction in asthma patients using FOT data, and interestingly, the simple k-nearest neighbor classifier achieved the best performance [67].

Telemedicine

Telemedicine originally emerged as a way to treat and manage patients who were located in remote places, far away from local health

facilities or in areas with shortages of medical professionals. As the affordability and accessibility to smartphones and sensors become widespread, telemedicine has become an important tool for the monitoring, self-management, and intervention of respiratory diseases. By monitoring clinical outcomes at an individual level, such technologies facilitate preventive and pre-emptive care remotely.

However, generating insights into the plethora of signals gathered by telemonitoring devices necessitates the development of ML algorithms. In the past, several researchers have applied predictive algorithms to monitor COPD and asthma exacerbations on data from telemonitoring devices [68]. Orchard et al. developed an ML model that considered daily reported clinical symptoms, physiological measurement of pulse and oxygen saturation, and medication to predict hospital readmissions and decision to start corticosteroids in COPD patients [69]. The ML model demonstrated a superior performance when compared to a traditional scoring model ($AUC = 0.74$ vs 0.60). Shah et al. carried out a similar study where they utilized a telemonitoring system consisting of a tablet computer with a customized application and a Bluetooth-enabled pulse oximeter [70]. The group trained a finite state ML model on the collected data to predict exacerbation events in COPD patients. An interesting mobile platform was developed by Chamberlain et al. [71], which consisted of an electronic stethoscope, peak flow meter, and a patient questionnaire. An ML algorithm was utilized to screen patients for COPD or asthma based on this data. ML methods such as naïve Bayes classifiers and support vector machines have been applied to home peak expiratory flow measurements and symptom scores to predict exacerbations a week early in adults [72] and children [73].

Miscellaneous

In this section, we highlight several applications of AI in miscellaneous applications of respiratory medicine.

Sleep Monitoring

Sleep monitoring or polysomnography (PSG) is a type of test to diagnose sleep disorders like obstructive sleep apnea [74]. Traditionally, PSG is performed in a sleep laboratory in which different sleep variables are monitored continuously overnight. One of the most cumbersome processes in polysomnography data is manual scoring of different sleep stages. To tackle this, many researchers have applied ML methods on PSG signals to automatically classify sleep stages [75–77]. Alloca et al. developed an automated sleep-stage classification program called “Somnivore,” which achieved a high F1 score (0.84–0.90) with manual visual scoring in humans, rodents, and pigeon polysomnography data [75]. Nikkinen et al. developed an DNN that accurately determined the oxygen desaturation index (ODI) and apnea–hypopnoea index (AHI) using only the oxygen saturation signal as input [78].

Breath Analysis

Breath analysis offers an excellent potential to phenotype respiratory disorders because exhaled breath contains a mixture of gases and traces of many volatile organic compounds (VOCs) that emanate from the respiratory tract itself. Several techniques exist to measure VOCs in the breath, such as gas chromatography–mass spectroscopy, electronic nose, and chemical sensors, each of which requires advanced pattern recognition methods to identify abnormal signatures in measured VOCs [79].

One of the generally accepted exhaled biomarkers is nitric oxide (NO), of which increased levels are associated with pulmonary inflammation in asthma patients. Machine learning models utilizing exhaled NO data have been used to discriminate between asthma and healthy individuals [80], to monitor asthma control in children [81], and to phenotype severe asthma [82]. ML methods on VOC data have been used to discriminate COPD and healthy individuals [83] and to detect lung cancer [84]. One study reported ten

new COPD-related VOCs, which was discovered using a data-mining approach [85].

Lung Sound Analysis

Computerized lung sound analysis involves discriminating between normal and adventitious lung sounds obtained during auscultation. Although ML has become a standard method for classifying adventitious sounds, these sound events are intermittent and highly variable from one person to another, presenting a challenge in generalizing these algorithms to a general population [86]. In the past, ML approaches have been applied to classify adventitious sounds associated with asthma [87], COPD [88], and ILD [89] and to detect common respiratory disorders in children using cough sounds [90]. Bardou et al. reported that DNNs outperformed traditional ML techniques in the classification of lung sounds into seven categories (normal, coarse crackle, fine crackle, monophonic wheeze, polyphonic wheeze, squawk, and stridor) [91].

Conclusions and Future Perspectives

We summarize our observations in Table 1. The application of artificial intelligence in respiratory medicine has created a very optimistic trend with new discoveries being reported at a breathtaking rate. DNNs are emerging as a key tool for developing imaging biomarkers for diagnosis, prognosis, and prediction of response to therapy. However, a major limitation for such computational approaches is the lack of large training datasets, which is characteristic of the medical field. An effective solution to this problem is transfer learning, an approach that entails pre-training a DNN on large nonmedical image datasets and then fine-tuning it in on an outcome specific dataset. Research has shown that transfer learning can outperform state-of-the-art ML models with manually engineered features in medical imaging [92].

There is also a need to develop open-source platforms to benchmark and bolster further deep learning research in respiratory medicine. The recent formation of open-source imaging consortium, a collaboration between academia and

Table 1 An overview of the applications of artificial intelligence across different modalities of respiratory medicine

Modality	Application	Main outcomes
Chest imaging	Plain chest X-ray	Common respiratory disease diagnosis [15–21], triaging [22], COVID-19 [23, 24]
	Computed tomography	Detecting COPD [26, 27], ILD [29, 30], pulmonary embolism [31, 32], identifying nodule malignancy [35–37], tumor prognosis [38, 39], COVID-19 [40–42]
	Histopathology	Tumor classification [44, 45], tumor growth pattern [46, 47]
	Bronchoscopy	Lung cancer diagnosis and classification [50, 51]
	Point-of-care ultrasound	Detecting B-line artifacts [53], COVID-19 [54]
Lung function testing	Spirometry	Obstructive airway disease differentiation and phenotyping [60, 61]
	Pulmonary function testing (spirometry, body plethysmography, and diffusion capacity)	Common respiratory disease diagnosis [58, 59]
	Forced oscillation test	Obstructive airway disease differentiation and phenotypes [64–67]
Telemedicine	Telemonitoring diagnostics	Predicting exacerbations in COPD and asthma [68–73]
Miscellaneous	Sleep monitoring	Scoring sleep stages
	Breath analysis of volatile organic compounds	Obstructive airway disease differentiation and phenotyping [64–67]
	Lung sound analysis	Classifying pathological lung sounds [86–91]

industry to develop imaging biomarkers for idiopathic pulmonary fibrosis (IPF), ILD, and other respiratory diseases using AI, is definitely a welcome sign [93]. There remains an enormous potential for DNNs to embrace domains outside of imaging, such as PFTs, FOTs, biosignal monitoring, etc. Since these applications involve the collection of raw data characterized by high dimensionality, DNNs may be employed to find specific patterns. The only bottleneck is the lack of sufficient sample sizes for training coupled with the diversity of the input data type that hinders transfer learning approaches. Finally, large clinical databases of multicenter randomized controlled intervention trials are another underexplored domain. DNNs on these detailed datasets carry the potential to predict treatment effects for individual patients. At this stage, the raw data of commercial datasets are highly protected and only available for internal use [94]. Merging raw data of similar studies for AI approaches may boost the field of personalized treatment.

We believe that the future has a lot in store for applications of AI. The recent trend of smartphones serving as general practitioners is only poised to grow with advances in Chabot capabilities that rely on natural language processing (NLP) [95]. This will make healthcare more accessible, especially in low-income countries, while bringing down overall costs [96]. Advances in NLP will also be deployed to extract clinical insights from the vast pool of unstructured electronic medical records (EMR), such as documents or notes [97]. The final frontier will lie in leveraging extremely large sets of data (or “big data”) that would integrate clinical, physiological, imaging, genomic sequence, and sensor data. DNNs would be used to explore quantitative biomarkers in big data and to develop personalized treatment and management strategies [98].

A criticism that is often directed at DNNs, and any other form of AI, is their lack of explainability, which can be defined as an ability to provide reasons for its output. Explainability is critical for gaining trust, especially if a physician plans to take action based on the prediction of a DNN. An explanation must be interpretable; therefore, revealing the mathematical operations of a DNN is hardly interpretable to the user. Several techniques currently exist today

that can generate explanations by estimating how the input features or different regions within an input image contributed to the output [99]. This may be useful in cases where the user could overrule the decision of a DNN by inspecting its faulty or biased explanations. However, there remains another technical impediment, that is, these systems are not sufficiently interrogated for robustness. It has been shown that DNNs can break down with small perturbations in input data [100]. For example, a DNN model trained on imaging data from the latest machine at an advanced care facility may malfunction at a rural hospital with older machines. However, this may not be the case for a human radiologist who can easily interpret data from both machines.

Lastly, we would like to highlight an “inconvenient truth” that while many AI algorithms have produced remarkable success in carefully selected cohorts (“laboratory conditions”), they are not yet ready to be deployed at the frontiers of clinical practice (“in the wild”) [101]. First, we should also consider updating our clinical guidelines and standards of care by incorporating AI, when it is evidence based, as it is done in the entire field of medicine [102]. Second, we would like to point out that the fragmented nature of our healthcare system, determined by a complex web of political, social, and commercial interests, provides very little incentive for AI applications. Thorny issues relating to data ownership, sovereignty, and transfer need to be resolved to allow AI innovations. At present, while the physician retains the final decision-making power, questions on liability arises as AI systems become more powerful and autonomous. Finally, the ethical implications of employing AI are poorly understood. AI models are prone not only to biases of the researchers or the institutions funding their development but also to spurious associations relating to gender or ethnicity in the datasets [103].

References

1. Russell SJ, Norvig P. Artificial Intelligence: a modern approach [Internet]. Neurocomputing. 1995;9: 215–218. <http://portal.acm.org/citation.cfm?id=773294>

2. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–44.
3. Aikins JS, Kunz JC, Shortliffe EH, Fallat RJ. PUFF: an expert system for interpretation of pulmonary function data. *Comput Biomed Res*. 1983;16(3):199–208.
4. van Ginneken B. Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning. *Radiol Phys Technol*. 2017;10(1):23–32.
5. Gonem S, Janssens W, Das N, Topalovic M. Applications of artificial intelligence and machine learning in respiratory medicine. *Thorax* [Internet]. 2020;75(8):695 LP–701. <http://thorax.bmjjournals.org/content/75/8/695.abstract>
6. Lodwick GS. Computer-aided diagnosis in radiology: a research plan. *Investig Radiol*. 1966;1(1):72–80.
7. Lodwick GS, Keats TE, Dorst JP. The coding of roentgen images for computer analysis as applied to lung cancer. *Radiology*. 1963;81(2):185–200.
8. Toriwaki J-I, Suenaga Y, Negoro T, Fukumura T. Pattern recognition of chest X-ray images. *Comput Graph Image Process* [Internet]. 1973;2(3):252–71. <http://www.sciencedirect.com/science/article/pii/0146664X73900051>
9. Kruger RP. A survey of computer processing of chest radiographs. In: Fifth international conference on information processing on medical imaging [Internet]. Nashville; 1977. p. 1689–99. <https://www.osti.gov/biblio/7214513>
10. Di Cataldo S, Ficarra E. Mining textural knowledge in biological images: applications, methods and trends. *Comput Struct Biotechnol J* [Internet]. 2017;15:56–67. <https://doi.org/10.1016/j.csbj.2016.11.002>.
11. Srinivasan G, Shobha G. Statistical texture analysis. *Proc world Acad* ... [Internet]. 2008;36 (December):1264–9. <http://staff.fh-hagenberg.at/wbackfri/Teaching/FBA/Uebungen/UE07charRecog/StatTextAnalysisSrinivasan08.pdf>
12. Sørensen L, Shaker SB, De Bruijne M. Quantitative analysis of pulmonary emphysema using local binary patterns. *IEEE Trans Med Imaging*. 2010;29(2):559–69.
13. Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Comput Med Imaging Graph*. 2007;31(4–5):198–211.
14. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou & K. Q. Weinberger (ed.), *Adv Neural Inf Process Syst*. 2012;1097–1105. Curran Associates, Inc.
15. Pasà F, Golkov V, Pfeiffer F, Cremers D, Pfeiffer F, et al. Efficient deep network architectures for fast chest x-ray tuberculosis screening and visualization. *Sci Rep*. 2019;9:6268. <https://doi.org/10.1038/s41598-019-42557-4>.
16. Qin ZZ, Sander MS, Rai B, et al. Using artificial intelligence to read chest radiographs for tuberculosis detection: a multi-site evaluation of the diagnostic accuracy of three deep learning systems. *Sci Rep*. 2019;9:15000. <https://doi.org/10.1038/s41598-019-51503-3>.
17. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning. 2017;3–9. <http://arxiv.org/abs/1711.05225>
18. Nam JG, Park S, Hwang EJ, et al. Development and validation of deep learning-based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. *Radiology*. 2019;290(1):218–228.
19. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med*. 2018;15 (11):e1002686. <https://doi.org/10.1371/journal.pmed.1002686>.
20. Hwang EJ, Park S, Jin KN, et al. Development and validation of a deep learning-based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA Netw Open*. 2019;2(3):e191095. <https://doi.org/10.1001/jamanetworkopen.2019.1095>.
21. Hwang EJ, Nam JG, Lim WH, et al. Deep learning for chest radiograph diagnosis in the emergency department. *Radiology*. 2019;293(3):573–580.
22. Annarumma M, Withey SJ, Bakewell RJ, Pesce E, Goh V, Montana G. Automated triaging of adult chest radiographs with deep artificial neural networks. *Radiology*. 2019;291(1):196–202.
23. Cohen JP, Dao L, Roth K, et al. Predicting COVID-19 pneumonia severity on chest x-ray with deep learning. *Cureus*. 2020;12(7):e9448. <https://doi.org/10.7759/cureus.9448>.
24. Minaee S, Kafieh R, Sonka M, Yazdani S, Soufi GJ. Deep-COVID: predicting COVID-19 from chest x-ray images using deep transfer learning. *Med Image Anal*. 2020;65:101794. <https://doi.org/10.1016/j.media.2020.101794>. PMID: 32781377; PMCID: PMC7372265.
25. Wynants L, Van Calster B, Collins G S, Riley R D, Heinze G, Schuit E et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ*. 2020;369:m1328. <https://doi.org/10.1136/bmj.m1328>.
26. Gonzalez G, Ash SY, Vegas-Sánchez-Ferrero G, et al. Disease staging and prognosis in smokers using deep learning in chest computed tomography. *Am J Respir Crit Care Med*. 2018;197(2):193–203.
27. Humphries SM, Notary AM, Centeno JP, et al. Deep learning enables automatic classification of emphysema pattern at CT. *Radiology*. 2020;294(2):434–444.
28. Das N, Topalovic M, Janssens W. Artificial intelligence in diagnosis of obstructive lung disease: current status and future potential. *Curr Opin Pulm Med* [Internet]. 2018;24(2). https://journals.lww.com/copulmonarymedicine/Fulltext/2018/03000/Artificial_intelligence_in_diagnosis_of.4.aspx

29. Walsh SLF, Calandriello L, Silva M, Sverzellati N. Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study. *Lancet Respir Med.* 2018;6(11):837–845.
30. Bermejo-Peláez D, Ash SY, Washko GR, et al. Classification of interstitial lung abnormality patterns with an ensemble of deep convolutional neural networks. *Sci Rep.* 2020;10:338. <https://doi.org/10.1038/s41598-019-56989-5>.
31. Tajbakhsh N, Shin JY, Gurudu SR, et al. Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE Trans Med Imaging.* 2016;35(5):1299–312.
32. Liu W, Liu M, Guo X, et al. Evaluation of acute pulmonary embolism and clot burden on CTPA with deep learning. *Eur Radiol.* 2020;30(6):3567–3575.
33. Delzell DAP, Magnuson S, Peter T, Smith M, Smith BJ. Machine learning and feature selection methods for disease classification with application to lung cancer screening image data. *Front Oncol.* 2019;9:1393. <https://doi.org/10.3389/fonc.2019.01393>. Erratum in: *Front Oncol.* 2020 Jun 05;10:866. PMID: 31921650; PMCID: PMC6917601.
34. Tu SJ, Wang CW, Pan KT, Wu YC, Wu CT. Localized thin-section CT with radiomics feature extraction and machine learning to classify early-detected pulmonary nodules from lung cancer screening. *Phys Med Biol.* 2018;63(6):065005. <https://doi.org/10.1088/1361-6560/aaafab>. PMID: 29446758.
35. Ardila D, Kiraly AP, Bharadwaj S, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med.* 2019;25:954–961.
36. Baldwin DR, Gustafson J, Pickup L, et al. External validation of a convolutional neural network artificial intelligence tool to predict malignancy in pulmonary nodules. *Thorax.* 2020;75:306–312.
37. Shen S, Han SX, Aberle DR, Bui AA, Hsu W. An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. *Expert Syst Appl.* 2019;128:84–95.
38. Hosny A, Parmar C, Coroller TP, Grossmann P, Zelenik R, Kumar A, Bussink J, Gillies RJ, Mak RH, Aerts HJWL. Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. *PLoS Med.* 2018;15(11):e1002711. <https://doi.org/10.1371/journal.pmed.1002711>. PMID: 30500819; PMCID: PMC6269088.
39. Wang S, Shi J, Ye Z, Dong D, Yu D, Zhou M, Liu Y, Gevaert O, Wang K, Zhu Y, Zhou H, Liu Z, Tian J. Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning. *Eur Respir J.* 2019;53(3):1800986. <https://doi.org/10.1183/13993003.00986-2018>. PMID: 30635290; PMCID: PMC6437603.
40. Li L, Qin L, Xu Z, et al. Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy. *Radiology* [Internet]. 2020;296(2):E65–71. <http://europemc.org/abstract/MED/32191588>
41. Wang S, Zha Y, Li W, Wu Q, Li X, Niu M, Wang M, Qiu X, Li H, Yu H, Gong W, Bai Y, Li L, Zhu Y, Wang L, Tian J. A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis. *Eur Respir J.* 2020;56(2):2000775. <https://doi.org/10.1183/13993003.00775-2020>. PMID: 32444412; PMCID: PMC7243395.
42. Huang L, Han R, Ai T, et al. Serial quantitative chest CT assessment of COVID-19: a deep-learning approach. *Radiol Cardiothorac Imaging.* 2020;2(2):e200075. Published 2020 Mar 30. <https://doi.org/10.1148/rct.2020200075>.
43. Shi H, Han X, Jiang N, et al. Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study. *Lancet Infect Dis.* 2020;20:425–434.
44. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med.* 2018;24:1559–1567.
45. Sha L, Osinski BL, Ho IY, Tan TL, Willis C, Weiss H, Beaubier N, Mahon BM, Taxter TJ, Yip SSF. Multi-field-of-view deep learning model predicts nonsmall cell lung cancer programmed death-ligand 1 status from whole-slide hematoxylin and eosin images. *J Pathol Inform.* 2019;10:24. https://doi.org/10.4103/jpi.jpi_24_19. PMID: 31523482; PMCID: PMC6669997.
46. Gertych A, Swiderska-Chadaj Z, Ma Z, et al. Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides. *Sci Rep.* 2019;9:1483. <https://doi.org/10.1038/s41598-018-37638-9>.
47. Wei JW, Tafe LJ, Linnik YA, et al. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Sci Rep.* 2019;9:3358. <https://doi.org/10.1038/s41598-019-40041-7>.
48. Pham HHN, Futakuchi M, Bychkov A, Furukawa T, Kuroda K, Fukuoka J. Detection of lung cancer lymph node metastases from whole-slide histopathologic images using a two-step deep learning approach. *Am J Pathol.* 2019;189(12):2428–2439.
49. Courtiol P, Maussion C, Moarii M, et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat Med.* 2019;25:1519–1525.
50. Tan T, Li Z, Liu H, et al. Optimize transfer learning for lung diseases in bronchoscopy using a new concept: sequential fine-tuning. *IEEE J Transl Eng Heal Med.* 2018;6:1800808. Published 2018 Aug 16. <https://doi.org/10.1109/JTEHM.2018.2865787>.
51. Feng PH, Chen TT, Lin YT, Chiang SY, Lo CM. Classification of lung cancer subtypes based on autofluorescence bronchoscopic pattern recognition: a preliminary study. *Comput Methods Prog Biomed.* 2018;163:33–38.

52. Melgarejo S, Schaub A, DVEN. Point of care ultrasound: an overview [Internet]. American College of Cardiology. <https://www.acc.org/latest-in-cardiology/articles/2017/10/31/09/57/point-of-care-ultrasound>
53. Cristiana B, Grzegorz T, Seungsoo K, et al. Automated lung ultrasound B-line assessment using a deep learning algorithm. *IEEE Trans Ultrason Ferroelectr Freq Control*. 2020;67:2312.
54. Roy S, Menapace W, Oei S, et al. Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound. *IEEE Trans Med Imaging*. 2020;39(8):2676–2687.
55. GINA. Diagnosis of diseases of chronic airflow limitation: Asthma, COPD and Asthma-COPD Overlap Syndrome (ACOS). Global Initiative for Chronic Obstructive Lung Disease; 2015.
56. Pellegrino R, Viegi G, Brusasco V, et al. Interpretative strategies for lung function tests. *Eur Respir J*. 2005;26(5):948–68.
57. Snow MG, Fallat RJ, Tyler WR, Hsu SP. Pulmonary consult: concept to application of an expert system. *J Clin Eng* [Internet]. 1988;13(3):201–5. http://journals.lww.com/jcejournal/Fulltext/1988/05000/Pulmonary_Consult_Concept_To_Application_Of_An.10.aspx
58. Topalovic M, Das N, Burgel PR, Daenen M, Derom E, Haenebalcke C, Janssen R, Kerstjens HAM, Liistro G, Louis R, Ninane V, Pison C, Schlesser M, Vercauter P, Vogelmeier CF, Wouters E, Wynants J, Janssens W. Artificial intelligence outperforms pulmonologists in the interpretation of pulmonary function tests. *Eur Respir J*. 2019;53(4):1801660. <https://doi.org/10.1183/13993003.01660-2018>. PMID: 30765505.
59. Le-Dong N-N, Hua-Huy T, Nguyen-Ngoc H-M, Dinh-Xuan A-T. Applying machine learning and pulmonary function data to detect interstitial lung disease in systemic sclerosis. *Eur Respir J*. 2017;50(suppl 61): OA3438. <https://doi.org/10.1183/1393003.congress-2017.OA3438>.
60. Bodduluri S, Nakhmani A, Reinhardt JM, et al. Deep neural network analyses of spirometry for structural phenotyping of chronic obstructive pulmonary disease. *JCI Insight*. 2020;5(13):e132781. Published 2020 Jul 9. <https://doi.org/10.1172/jci.insight.132781>.
61. Kaplan A, Cao H, Fitzgerald JM, et al. Asthma/COPD Differentiation Classification (AC/DC): Machine Learning to Aid Physicians in Diagnosing Asthma, COPD and Asthma-COPD Overlap (ACO). In: D22 COMORBIDITIES IN PEOPLE WITH COPD [Internet]. American Thoracic Society; 2020. p. A6285–A6285. (American Thoracic Society International Conference Abstracts). https://doi.org/10.1164/ajrccm-conference.2020.201.1_MeetingAbstracts.A6285
62. Das N, Verstraete K, Stanojevic S, Topalovic M, Aerts J-M, Janssens W. Deep learning algorithm helps to standardise ATS/ERS spirometric acceptability and usability criteria. *Eur Respir J* [Internet]. 2020;2000603. <http://erj.ersjournals.com/content/early/2020/06/08/13993003.00603-2020.abstract>
63. Brashier B, Salvi S. Measuring lung function using sound waves: role of the forced oscillation technique and impulse oscillometry system. *Breathe*. 2015;11(1): 57–65.
64. Amaral JLM, Lopes AJ, Jansen JM, Faria ACD, Melo PL. Machine learning algorithms and forced oscillation measurements applied to the automatic identification of chronic obstructive pulmonary disease. *Comput Methods Programs Biomed* [Internet]. 2012;105(3): 183–93. <https://doi.org/10.1016/j.cmpb.2011.09.009>.
65. Amaral JLM, Lopes AJ, Faria ACD, Melo PL. Machine learning algorithms and forced oscillation measurements to categorise the airway obstruction severity in chronic obstructive pulmonary disease. *Comput Methods Prog Biomed*. 2015;118(2):186–97.
66. Amaral JLM, Lopes AJ, Jansen JM, Faria ACD, Melo PL. An improved method of early diagnosis of smoking-induced respiratory changes using machine learning algorithms. *Comput Methods Programs Biomed* [Internet]. 2013;112(3):441–54. <https://doi.org/10.1016/j.cmpb.2013.08.004>.
67. Amaral JLM, Lopes AJ, Veiga J, Faria ACD, Melo PL. High-accuracy detection of airway obstruction in asthma using machine learning algorithms and forced oscillation measurements. *Comput Methods Prog Biomed*. 2017;144:113–25.
68. Sanchez-Morillo D, Fernandez-Granero MA, Leon-Jimenez A. Use of predictive algorithms in-home monitoring of chronic obstructive pulmonary disease and asthma: a systematic review. *Chron Respir Dis* [Internet]. 2016;13(3):264–83. <http://www.sagepub.com/journalsProdDesc.nav?prodId=Journal201805%5Cn http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed14&NEWS=N&AN=20160583437>
69. Orchard P, Agakova A, Pinnock H, Burton CD, Sarran C, Agakov F, McKinstry B. Improving prediction of risk of hospital admission in chronic obstructive pulmonary disease: application of machine learning to telemonitoring data. *J Med Internet Res*. 2018;20(9):e263. <https://doi.org/10.2196/jmir.9227>. PMID: 30249589; PMCID: PMC6231768.
70. Shah SA, Velardo C, Farmer A, Tarassenko L. Exacerbations in chronic obstructive pulmonary disease: identification and prediction using a digital health system. *J Med Internet Res*. 2017;19(3):e69.
71. Chamberlain DB, Kodgule R, Fletcher RR. A mobile platform for automated screening of asthma and chronic obstructive pulmonary disease. In: Proceedings of the annual international conference of the IEEE engineering in medicine and biology society, EMBS. 2016.
72. Finkelstein J, Jeong IC. Machine learning approaches to personalize early prediction of asthma exacerbations. *Ann N Y Acad Sci*. 2017;1387(1):153–165.

73. Luo G, Stone BL, Fassl B, et al. Predicting asthma control deterioration in children. *BMC Med Inform Decis Mak.* 2015;15(8). <https://doi.org/10.1186/s12911-015-0208-9>.
74. Medical Advisory Secretariat. Polysomnography in patients with obstructive sleep apnea: an evidence-based analysis. *Ont Health Technol Assess Ser.* 2006;6(13):1–38.
75. Allocca G, Ma S, Martelli D, et al. Validation of ‘Somnivore’, a machine learning algorithm for automated scoring and analysis of polysomnography data. *Front Neurosci.* 2019;13:207. Published 2019 Mar 18. <https://doi.org/10.3389/fnins.2019.00207>.
76. Mousavi S, Afghah F, Acharya UR. SleepEEGNet: automated sleep stage scoring with sequence to sequence deep learning approach. *PLoS One.* 2019;14(5):e0216456. Published 2019 May 7. <https://doi.org/10.1371/journal.pone.0216456>.
77. Yildirim O, Baloglu UB, Acharya UR. A deep learning model for automated sleep stages classification using PSG signals. *Int J Environ Res Public Health.* 2019;16(4):599. Published 2019 Feb 19. <https://doi.org/10.3390/ijerph16040599>.
78. Nikkonen S, Afara IO, Leppänen T, Töyräs J. Artificial neural network analysis of the oxygen saturation signal enables accurate diagnostics of sleep apnea. *Sci Rep.* 2019;9(1):13200. <https://doi.org/10.1038/s41598-019-49330-7>. Erratum in: *Sci Rep.* 2020 Mar 13;10(1):4977. PMID: 31519927; PMCID: PMC6744469.
79. Christiansen A, Davidsen JR, Titlestad I, et al. A systematic review of breath analysis and detection of volatile organic compounds in COPD. *J Breath Res [Internet].* 2016;10(3):034002. <http://stacks.iop.org/1752-7163/10/i=3/a=034002?key=crossref.92312f7e0ef6e081d8e4d215ece42a9e>
80. Montuschi P, Santonico M, Mondino C, et al. Diagnostic performance of an electronic nose, fractional exhaled nitric oxide, and lung function testing in asthma. *Chest.* 2010;137(4):790–6.
81. Pifferi M, Bush A, Pioggia G, et al. Monitoring asthma control in children with allergies by soft computing of lung function and exhaled nitric oxide. *Chest.* 2011;139(2):319–27.
82. Wu W, Bleeker E, Moore W, et al. Unsupervised phenotyping of severe asthma research program participants using expanded lung data. *J Allergy Clin Immunol.* 2014;133(5):1280–8.
83. Phillips CO, Syed Y, Parthalán NM, Zwigelaar R, Claypole TC, Lewis KE. Machine learning methods on exhaled volatile organic compounds for distinguishing COPD patients from healthy controls. *J Breath Res [Internet].* 2012;6(3):036003. <http://stacks.iop.org/1752-7163/6/i=3/a=036003?key=crossref.504b3410c7803d5c45812740e3fb1868>
84. Huang CH, Zeng C, Wang YC, Peng HY, Lin CS, Chang CJ, Yang HY. A study of diagnostic accuracy using a chemical sensor array and a machine learning technique to detect lung cancer. *Sensors (Basel).* 2018;18(9):2845. <https://doi.org/10.3390/s18092845>. PMID: 30154385; PMCID: PMC6164114.
85. Basanta M, Ibrahim B, Dockry R, et al. Exhaled volatile organic compounds for phenotyping chronic obstructive pulmonary disease: a cross-sectional study. *Respir Res [Internet].* 2012;13(1):72. <http://respiratory-research.com/content/13/1/72>
86. Pramono RXA, Bowyer S, Rodriguez-Villegas E. Automatic adventitious respiratory sound analysis: a systematic review. *PLoS One.* 2017;12(5):e0177926.
87. Islam MA, Bandyopadhyaya I, Bhattacharyya P, Saha G. Multichannel lung sound analysis for asthma detection. *Comput Methods Programs Biomed.* 2018;159:111–123.
88. Jácome C, Marques A. Computerized respiratory sounds in patients with COPD: a systematic review. *COPD [Internet].* 2014;2555(November):1–9. <http://www.ncbi.nlm.nih.gov/pubmed/24914587>
89. Flietstra B, Markuzon N, Vyshedskiy A, et al. Automated analysis of crackles in patients with interstitial pulmonary fibrosis. *Pulm Med.* 2011;2011:590506.
90. Porter P, Abeyratne U, Swarnkar V, et al. A prospective multicentre study testing the diagnostic accuracy of an automated cough sound centred analytic system for the identification of common respiratory disorders in children. *Respir Res.* 2019;20:81. <https://doi.org/10.1186/s12931-019-1046-6>.
91. Bardou D, Zhang K, Ahmad SM. Lung sounds classification using convolutional neural networks. *Artif Intell Med.* 2018;88:58–69.
92. Ravishankar H, Sudhakar P, Venkataramani R, et al. Understanding the mechanisms of deep transfer learning for medical images. In: Lecture notes in computer science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2016.
93. Open Source Imaging Consortium (OSIC). Interstitial Lung Disease (ILD) Experts and Advocates Announce Formation of Open Source Imaging Consortium (OSIC) [Internet]. 2019 [cited 2019 Nov 10]. <https://www.osicild.org/press/osicild-attends-ucl>
94. Naudet F, Sakarovich C, Janiaud P, Cristea I, Fanelli D, Moher D, Ioannidis JPA. Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: survey of studies published in The BMJ and PLOS Medicine. *BMJ.* 2018;360:k400. <https://doi.org/10.1136/bmj.k400>. PMID: 29440066; PMCID: PMC5809812.
95. Palanica A, Flaschner P, Thommandram A, Li M, Fossat Y. Physicians’ perceptions of chatbots in health care: cross-sectional web-based survey. *J Med Internet Res.* 2019;21(4):e12887. Published 2019 Apr 5. <https://doi.org/10.2196/12887>.
96. Wahl B, Cossy-Gantner A, Germann S, et al. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? *BMJ Glob Health.* 2018;3:e000798.

97. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR Med informatics* [Internet]. 2019;7(2):e12239. <https://www.ncbi.nlm.nih.gov/pubmed/31066697>
98. Viceconti M, Hunter P, Hose R. Big data, big knowledge: big data for personalized healthcare. *IEEE J Biomed Heal Informatics*. 2015;19(4):1209–15.
99. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: an overview of interpretability of machine learning. In: Proceedings – 2018 IEEE 5th international conference on data science and advanced analytics, DSAA 2018. 2019.
100. Heaven D. Why deep-learning AIs are so easy to fool. *Nature*. 2019;574(7777):163–166. <https://doi.org/10.1038/d41586-019-03013-5>. PMID: 31597977.
101. Panch T, Mattie H, Celi LA. The “inconvenient truth” about AI in healthcare. *NPJ Digit Med*. 2019;2:77.
102. Sung JJY, Stewart CL, Freedman B. Artificial intelligence in health care: preparing for the fifth Industrial Revolution. *Med J Aust* [Internet]. 2020. <https://doi.org/10.5694/mja2.50755>.
103. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf*. 2019;28:231.



Ching-Heng Lin and Chang-Fu Kuo

Contents

Introduction	774
Application of Artificial Intelligence in Rheumatology	775
Application in Electronic Health Records	776
Application in Genetic and Biomarker Data	779
Application in Medical Images	780
Application in Mixture Data	781
Future Perspectives and Challenges	782
Conclusions	782
References	782

Abstract

Rheumatology is a discipline to manage clinical conditions in joints, soft tissue, connective tissue, and autoimmune diseases. Rheumatoid arthritis, systemic lupus erythematosus and other connective tissue diseases, gout and metabolic bone diseases, osteoarthritis, and spondyloarthritis are well-known multi-organ rheumatic diseases. The application of artificial intelligence (AI) in rheumatology is relatively

rare; however, an increase in AI research in rheumatology is apparent, especially in the recent 2 years. AI has been applied in prediction of disease or outcome using electronic health records (such as for axial spondyloarthritis), deciphering of genetic or other high-dimensional biological data (such as for erosive arthritis, juvenile idiopathic arthritis), image classification (such as for anti-nuclear antibody, joint space narrowing, fractures), and miscellaneous AI applications in mixed types of data (such as for inflammatory synovial). This book chapter summarizes recent AI research in rheumatology and provides a perspective on the potential of AI in clinical or research aspects of rheumatology.

C.-H. Lin
Center for Artificial Intelligence in Medicine, Chang Gung Memorial Hospital, Taoyuan, Taiwan

C.-F. Kuo (✉)
Center for Artificial Intelligence in Medicine, Chang Gung Memorial Hospital, Taoyuan, Taiwan

Division of Rheumatology, Allergy, and Immunology,
Chang Gung Memorial Hospital, Taoyuan, Taiwan

Keywords

Rheumatology · Rheumatoid arthritis ·
Systemic lupus erythematosus ·

Osteoarthritis · Gout · Osteoporosis · Artificial intelligence · Machine learning · Deep learning

Introduction

Rheumatology is a unique discipline to manage patients with rheumatic diseases, medical conditions that primarily affect the bone, joint, and soft tissue, and connective diseases. Many of these diseases have autoimmunity in pathogenesis; therefore, newer therapeutics target specific immunological proteins to remedy aberrant immune function. The diagnosis of rheumatic diseases is not straightforward, often requiring multiple clinical tools to help rheumatologists decide. The application of artificial intelligence (AI) in rheumatology is still in infancy, but some research has documented the potential to transform the practice of rheumatology.

The major rheumatic diseases include rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), metabolic bone diseases (e.g., gout and osteoporosis), spondyloarthritis (e.g., ankylosing spondylitis and psoriatic arthritis), and other connective tissue diseases (e.g., systemic sclerosis). RA is an inflammatory polyarthritis primarily affecting small joints, which often causes disability if left untreated. Efforts to diagnose RA early include the use of *machine learning* (ML) to identify disease cases from electronic health records (EHRs) and assist the classification of RA. SLE is a prototypic autoimmune disease characterized by the presence of multiple autoantibodies and a myriad of clinical presentation. AI/ML has been used to identify or classify SLE patients based on genetic or immunophenotyping data. Gout is a urate deposition disease that often causes acute arthritis. The prediction of acute flare is a challenge to clinicians and the patients since the prophylactic medication is effective and can be safely administered prior to the flare. ML has been used to automate the identification of acute flare from clinical notes that can further accelerate the development of a predictive model. Spondyloarthritis is a group of inflammatory arthritis affecting axial (ankylosing spondylitis) or peripheral joints (psoriatic arthritis). The diagnosis of these conditions

heavily relies on bone and joint radiography. AI has been applied to assist in interpreting X-rays and tracing the radiographic progression of the diseases. Osteoporosis results from a decrease in bone mineral density, which reduces bone strength and increases the risk of fractures. Since many of the affected patients are asymptomatic, effective screening to identify patients at risk is essential, where AI algorithms are helpful.

AI is a set of algorithms and technologies used to imitate or augment human intelligence. Figure 1 depicts the research disciplines in AI that include, but are not limited to, planning, reasoning, knowledge representation perception, robotics, and machine learning. In machine learning, various algorithms are developed to detect patterns in sampled data for prediction or decision-making. Among these machine learning algorithms, the artificial neural network technique has received much attention in recent years due to its success in many applications (e.g., macular degeneration detection, self-driving car, or virtual assistant). The technical breakthroughs of neural network training and design become an emerging subdiscipline, deep learning. Different types of neurons and various layer connectivity styles form numerous deep learning models.

Machine learning, as an important subfield of AI, is characterized by its ability to identify clinically relevant patterns among an abundance of medical data by different learning strategies. Supervised learning is the most popular machine learning strategy where the model is trained on a labeled dataset to recognize the patterns associated with specific classes. This type of learning strategy is commonly used for classification and regression. To date, many of the machine learning studies in rheumatology have used supervised learning models such as support vector machine (SVM), logistic regression, random forest, or artificial neural network (ANN) to perform disease classification or detection. Supervised learning is often used in tasks with a clear reference label, such as the prediction of patient outcomes using electronic medical charts.

On the contrary, unsupervised learning is another learning strategy using unlabeled data to detect the underlying relationships or patterns of

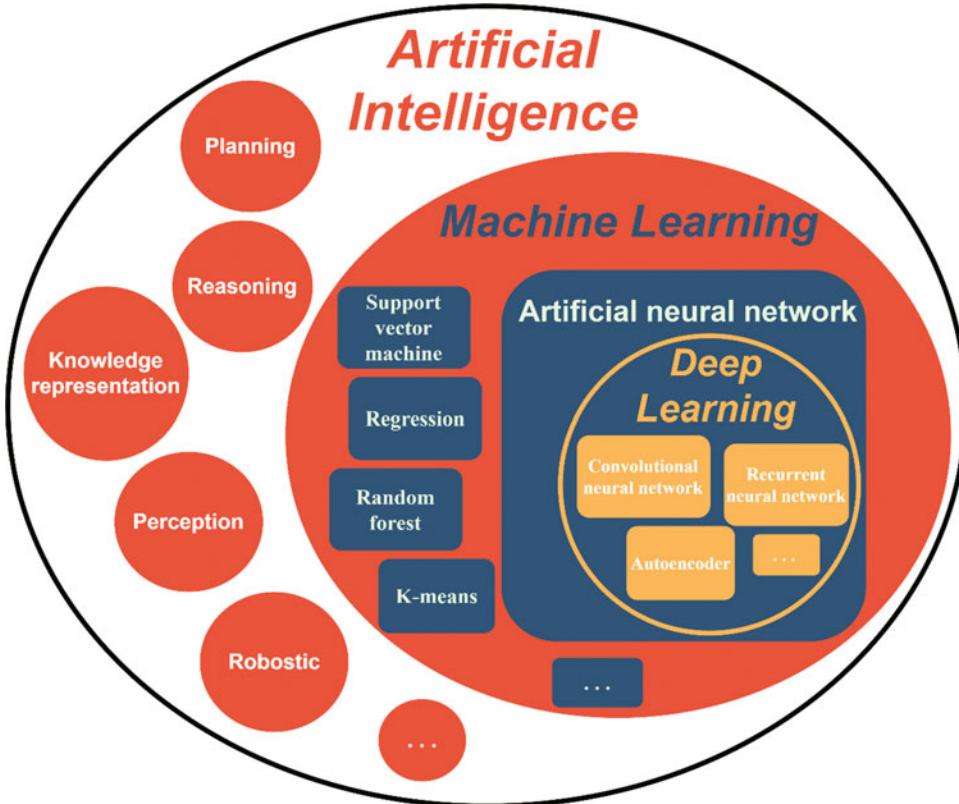


Fig. 1 The Venn diagram of different artificial intelligence disciplines

the dataset. The unsupervised learning algorithm can be a useful data analytics tool for clinicians to gain a better understanding of the medical data because it can reduce the dimensionality of the given data (e.g., t-SNE and autoencoder) or cluster similar data points (e.g., k-means and DBSCAN). These algorithms delineate the inherent data structure and provide insight into complex data such as genetic or high-dimensional flow cytometric data.

As a specific class of machine learning algorithms, deep learning has received much spotlight in recent years and has already been successfully used in many medical disciplines such as ophthalmology, radiology, and pathology. Deep neural network (DNN), a type of ANN with multiple designed layers, is an essential form of deep learning. In contrast to classic machine learning algorithms that still require feature engineering with prior knowledge, DNN automatically extracts features from the raw data by nonlinear

transformations of neurons (hidden layers). Convolutional neural network (CNN) is one of the prevalent DNN architectures. CNN has been mainly applied in medical imaging tasks. It recognizes feature patterns by the convolutional layer, learning different phenotypic signatures of a medical image. Recurrent neural network (RNN) is another famous DNN architecture. RNN is specialized for processing sequential data by allowing previous outputs to be used as inputs while having hidden states that make it more suitable for longitudinal data analysis.

Application of Artificial Intelligence in Rheumatology

AI is reshaping rheumatology in both basic research and clinical practice since it can perform tasks that usually would require medical expertise

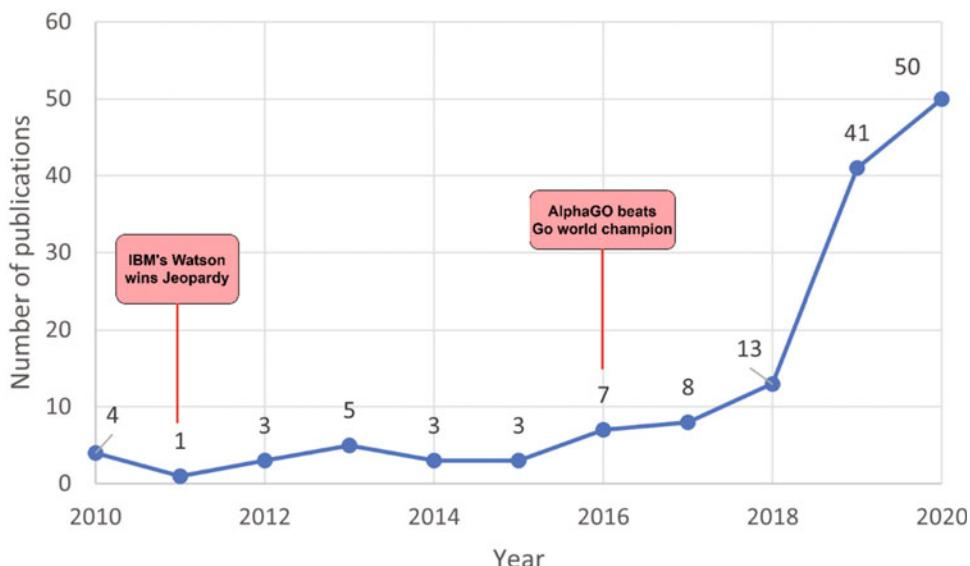


Fig. 2 Number of studies of artificial intelligence in rheumatology on PubMed (between 2010 and 9 November 2020). Search query: ((“rheumatology”[Title/Abstract] OR “rheumatic”[Title/Abstract] OR “rheumatoid”[Title/

Abstract]) AND (“machine learning”[Title/Abstract] OR “artificial intelligence”[Title/Abstract] OR “deep learning”[Title/Abstract]))

such as diagnosis, prognosis prediction, or medical image analysis. Research and clinical applications of AI, including machine learning and deep learning, in rheumatology, are rare, but the number and diversity of AI research in the field has been raised during the last decade (Fig. 2).

In the following paragraphs, we discussed recent publications in the field relevant to AI. We categorize the publications according to the data types: electronic health records, genetic data or biomarkers, medical images, and mixed types of data (listed in Table 1).

Application in Electronic Health Records

The EHRs can be a valuable tool for conducting research studies on rheumatology patients’ disease, treatment, and outcomes since EHRs contain a tremendous amount of patient information. AI models that trained from EHRs have been increasingly adopted as a disease detector or a predictor. For instance, Lötsch et al. [1] assemble machine learning algorithms to identify the predictors

associated with the development of persistent pain in RA. A Gaussian mixture model (GMM) was applied to identify the three distinct pain intensity subgroups (low, median, and high persistent). After the unsupervised clustering, a random forest was used to select predictive parameters from clinical data to train machine learning classifiers. Classifiers trained with selected parameters provided pain group assignment at a balanced accuracy of 70%. Identification of RA patients in primary care EHR is clinically essential for prompt treatment. In order to diagnose RA at an early stage, the eDRAM (early disease risk assessment) framework was proposed for finding important risk factors from a large-scale EHR database by nonnegative matrix factorization (NMF) and performing SVM for disease risk assessment [2]. The eDRAM was evaluated in a cohort of 1,007 RA-diagnosed patients and 921,192 control patients. The result shows that the eDRAM had an accuracy of 72% with 200 risk factor inputs. The other study applied the random forest model to identify the most informative predictors to detect RA patients from primary care EHRs [3]. The model identified eight predictors,

Table 1 Artificial intelligence studies in rheumatic diseases

Input type	Applications	AI methods	Optimal result	Ref.
Electronic health records	Identify the predictors associated with the development of persistent pain in rheumatoid arthritis	Gaussian mixture model	Balanced accuracy: 70%	[1]
	Diagnose rheumatoid arthritis at an early stage	Nonnegative matrix factorization, support vector machine	Accuracy: 72%	[2]
	Identify the most informative predictors to detect rheumatoid arthritis patients	Random forest model	Eight predictors were identified	[3]
	Identify systemic lupus erythematosus cases from electronic health record data	L1-regularized logistic regression	AUROC: 0.97	[4]
	Detect the occurrence of osteoarthritis from health behavior survey data	Deep neural network	AUROC: 0.768	[5]
	Mortality prediction	Random survival forests	Specificity: 0.8, sensitivity: 0.48	[6]
	Predict radiographic progression in patients with axial spondyloarthritis	Support vector machine	Average AUROC: 0.78	[7]
	Predict readmission risk to an outpatient rheumatology clinic	Random forest	AUROC: 0.653	[8]
	Forecast the state of disease activity of rheumatoid arthritis patients	Recurrent neural network	AUROC: 0.91	[9]
	Predict hospital readmission for lupus patients	Long short-term memory network	AUROC: 0.7	[10]
Genetic and biomarker data	Identify evidence of gout flares from the medical records	Natural language processing and support vector machine	Sensitivity, 82.1%; specificity, 91.5%	[11]
	Identify rheumatoid arthritis cases from medical records	Support vector machine	AUROC: 0.966	[12]
	Identify systemic lupus erythematosus cases from medical records	Penalized logistic regression	Specificity, 97%, sensitivity, 60%	[13]
	Identify key biomarkers to discriminate between healthy people and seropositive rheumatoid arthritis patients	Artificial neural network	Accuracy: 83.3%	[14]
	Identify biomarkers associated with ultrasonography-detected erosive arthritis	Logistic regression, decision trees	Six biomarkers were identified	[15]

(continued)

Table 1 (continued)

Input type	Applications	AI methods	Optimal result	Ref.
	Stratify patients with systemic lupus erythematosus based on gene expression signatures	k-means clustering, support vector machine	Accuracy: 88%	[19]
	Stratify patients with juvenile-onset systemic lupus erythematosus	Balanced random forest	Accuracy of 90.9%	[20]
	Distinguish methotrexate therapeutic nonresponders from responders	L2-regularized logistic regression	AUROC: 0.78	[21]
	Predict systemic lupus erythematosus disease activity	Random forest	Accuracy: 83%	[22]
Medical images	Detect and estimate geometric and texture features of synovium thickening and bone erosion	Support vector machine	Accuracy: 92.50%	[23]
	Rheumatoid arthritis image grading	Convolutional neural network	Accuracy: 92.50% and 95% in two datasets	[24]
	Osteoarthritis image severity grading	Assembled DenseNet	Sensitivity, 86.0%; specificity, 99.1%	[25]
	Scoring of rheumatoid arthritis activity from Doppler ultrasound images	Convolutional neural network	Accuracy: 75.0%	[26]
	Diagnose hip osteoarthritis from X-ray images	Convolutional neural network	Accuracy of 92.8%	[27]
	Diagnose knee osteoarthritis from X-ray images	Siamese convolutional neural network	AUROC: 0.93	[28]
	Detect cartilage lesions at MRI for osteoarthritis diagnosis	Convolutional neural network	AUROC: 0.91	[29]
	Distinguish between eroded and not eroded joints	Convolutional neural network	Balanced accuracy of 47.5%	[30]
	Quantify the difference between healthy and eroded image patches	U-net, Siamese convolutional neural network	–	[31]
	Detect low bone mineral density using plain hip radiographs	Deep texture encoding network	Sensitivity, 91.11%; specificity, 86.08%	[32]
Mixture data	Assess the radiographic bone texture in the distal metacarpal bone relevant to rheumatoid arthritis	Graph convolutional network, deep texture encoding network	AUROC: 0.68	[33]
	Detect radiographic osteoarthritis	Deep texture encoding network, ResNet18	Positive predictive value: 81.37% for deep-TEN, 87.46% for ResNet18	[34]
	Identify synovial subtype, and determine the synchrony between synovial histologic features and their genomic subtypes	k-means clustering, support vector machine	AUROC: 0.88 for high inflammatory subtype, 0.71 for high inflammatory subtype, 0.59 for mixed subtype	[35]
	Predict treatment responses in rheumatoid arthritis patients	Gaussian process regression	AUROC: 0.66	[36]
	Predict osteoarthritis progression	Multimodal machine learning	AUROC: 0.79	[37]

including diagnostic codes for RA, medication codes, and absence of alternative diagnoses. Murray et al. [4] trained an L1-regularized logistic regression with the EasyEnsemble method on a noisy labeled dataset to automatically identify SLE cases from EHR data. The final model achieved an AUROC of 0.97. In addition to image-based deep learning algorithms on automatically detecting osteoarthritis (OA), Lim et al. [5] used a DNN with scaled principal component analysis (PCA) to detect the occurrence of OA from health behavior survey data. Their experiments show that the DNN with scaled PCA resulted in 76.8% of AUROC and minimized the effort to generate features.

In terms of prognosis or outcome prediction, a mortality prediction model using the random survival forests (RSF) was developed by Lezcano-Valverde et al. [6]. The RSF model identified five mortality risk groups and has a specificity of 0.8 and a sensitivity of 0.48 in the validation cohort. Joo et al. [7] carried out a retrospective and hospital-based study to evaluate seven machine learning-based models on predicting radiographic progression in patients with axial spondyloarthritis (axSpA). The SVM was the top best-performing models (average AUROC was over 0.78) in their study. Prevention of readmission is an important task in RA patient management. Madrid-García et al. [8] developed a random forest model from departmental EHR data for predicting readmission risk to an outpatient rheumatology clinic between 2 and 12 months after discharge. The final model's performance was AUROC of 0.653 on a chronologically split test dataset. Additionally, to leverage longitudinal data in EHR, the RNN has been widely adopted in recent years. Norgeot et al. [9] assessed the ability of an RNN to forecast the state of disease activity of RA patients at their next clinical visit. The predictive model reached an AUROC of 0.91 in a test cohort of 116 patients. Another example is applying long short-term memory (LSTM), a type of RNN, to predict hospital readmission for lupus patients. The LSTM has a better performance compared to traditional ANN (0.7 AUROC vs. 0.66 AUROC) [10].

Most of the clinical information in EHRs is unstructured narrative; therefore, various studies explored the integration of natural language processing (NLP) and ML techniques to unleash the power of EHRs in rheumatology research. Zheng et al. [11] developed an automatic tool that combines NLP and ML algorithms for identifying evidence of gout flares from the EHRs. This tool was evaluated using over 590,000 clinical notes; it identified 18,869 clinical notes as gout flare positive with a sensitivity of 82.1% and a specificity of 91.5%. Carroll et al. [12] trained an SVM model based on a refined set of EHR features containing ICD-9 codes, medication exposures, and NLP-derived information for identifying RA cases. Their study demonstrated that compared to naïve collections of EHR features, using a refined set of features to train an SVM can achieve slightly better performance (AUROC of 0.966 on refined dataset vs. AUROC of 0.956 on naïve dataset) and required relatively fewer cases (50–100 training samples are required) on RA patient identification. Jorge et al. [13] used penalized logistic regression to identify SLE cases from EHRs. However, it is worth noting that, as a result, adding SLE-related concepts extracted from unstructured data via NLP did not improve the model's performance.

Application in Genetic and Biomarker Data

Much research effort has been focused on the discovery of biomarkers that enable early diagnosis of RA. Nowadays, artificial intelligence has become a powerful tool in biomarker discovery. Chocholova et al. [14] performed data mining using ANN on serum sample data to identified key biomarkers, which can discriminate between healthy people and seropositive RA patients with an accuracy of 83.3%. In another study, logistic regression and decision trees with the forward wrapper feature selection method were applied to a cohort of 120 SLE patients to identify biomarkers associated with ultrasonography-detected erosive arthritis. There were six features selected by the machine learning models,

including anti-CarP, ACPA, arthralgia, Jaccoud's arthropathy, anti-Sm, and neurological manifestations. Their further analysis reveals that ACPA and anti-CarP antibodies have a relatively higher importance in developing erosive damage [15].

Next-generation sequencing (NGS) and transcriptomics provide an unprecedented genome-wide view of gene expression in rheumatology research. Poppenberg et al. [16] trained four different machine learning models with transcriptomes from peripheral blood mononuclear cells (PBMCs) for predicting juvenile idiopathic arthritis stage. Random forest outperformed other models with an AUROC of 0.94 on an independent testing cohort of 14 samples. To predict psoriatic arthritis (PsA) risk among psoriasis patients by genetic signature, Patrick et al. [17] adopted machine-learning techniques for genetic architecture subtype prediction and risk assessment. The PsA prediction model achieved > 90% precision with a specificity of 100%. Supervised machine learning classifiers were also employed to aid clinical decision-making. For example, Franks et al. [18] conducted a study on identifying four intrinsic gene expression subsets of systemic sclerosis (SSc). They built multinomial elastic net (GLMnet), SVM, and random forest classifiers with gene expression data of 297 skin biopsy samples that merged from three independent cohorts. The external validation shows that GLMnet achieved the best accuracy of 85.4%. In a study of stratification of patients with SLE based on gene expression signatures derived from whole-blood transcriptomic data, machine learning methods including k-means clustering and SVM with error-correcting output code (ECOC) framework were performed as a new stratification scheme. The classifier was trained to learn the transcriptomic signatures of each cluster and accurately classify new patients with an accuracy of 88% [19]. Another study related to SLE stratification was conducted by Robinson et al. [20]. This retrospective study collected 67 patients with juvenile-onset SLE and 39 healthy controls. A balanced random forest (BRF) model was built for discriminating patients with juvenile-onset SLE from healthy

controls with a prediction accuracy of 90.9%. A highly predictive classifier of methotrexate (MTX) nonresponse was developed by Plant et al. [21] using L2-regularized logistic regression. This predictive model has an AUROC of 0.78 on distinguishing therapeutic nonresponders from responders using whole-blood transcript expression data. The high degree of SLE gene expression heterogeneity among patients and study cohorts is challenging SLE disease activity prediction. Kegerreis et al. [22] established a random forest classifier to integrate gene expression data from three SLE datasets and used it to predict SLE disease activity. The classifier achieved a peak classification accuracy of 83% under tenfold cross-validation.

Application in Medical Images

Medical imaging is essential in the diagnosis and staging of RA disease. There is a long history of classic machine learning methods in rheumatology imaging, and recently there have been many studies adopted deep learning approaches to solve more complex problems. The grading evaluation of metacarpophalangeal (MCP) RA ultrasonic images heavily relies on trained sonographers' expertise. Yang et al. [23] proposed a computer-aided grading method for detecting and estimating geometric and texture features of synovium thickening and bone erosion. They extracted quantitative feature metrics and texture features from the region of interest (ROI) of RA ultrasound images. The SVM classifier, trained with ROI feature descriptor, provides the highest accuracy of 92.50% on classifying four grading MCP RA ultrasonic images.

In contrast to feature engineering and classic machine learning, deep learning automatically learns texture and image context features from training data. Tang et al. [24] developed an automatic RA grading method using deep CNN architecture. The model takes gray-scale ultrasound images of MCP and proximal-interphalangeal (PIP) joints as inputs and then outputs the corresponding RA grading results. The deep CNN-based RA grading model achieves

satisfactory accuracy (90.3% in MCP dataset and 95% in PIP dataset, respectively), comparable to RA experts in four-grade classification.

Radiographic scoring systems play an important role in the evaluation of the progression and treatment of RA. Kellgren-Lawrence (KL) grading system is widely used for clinical assessment and diagnosis of OA. Norman et al. [25] grouped KL grades into four degrees of OA severity (KL 0–1 are no OA; KL 2–4 represent mild, moderate, and severe OA, respectively) and developed an assembled DenseNet with over 4,000 bilateral PA fixed-flexion knee radiographs to make OA assessments. This assembled DenseNet's testing sensitivity and specificity of no OA were 83.7% and 86.1%; mild OA was 70.2% and 83.8%; moderate OA were 68.9% and 97.1%, and severe OA was 83.7% and 99.1%, respectively. In addition, the saliency map analysis showed that the assembled DenseNet could correctly identify relevant features within a radiographic image that are used to make OA assessments, such as joint space narrowing and osteophytes. In another example of AI application in the radiographic scoring system, a study conducted by Andersen et al. [26] showed that CNN could be used in the scoring of RA activity from Doppler ultrasound images according to the OMERACT-EULAR Synovitis Scoring (OEES) system. A total of 1,694 Doppler ultrasound images were used in this study. The CNN model achieved an average per class (OEES scores of 0–3) accuracy of 75.0%.

With the advantage of machine learning, various studies developed the computer-aided diagnosis tool of OA in medical imaging in rheumatology. A deep CNN was trained and tested on 420 hip X-ray images to diagnose hip OA. The CNN model performance is comparable to an attending physician with 10 years of experience with an accuracy of 92.8% [27]. Knee OA is another common musculoskeletal disorder in RA patients. Tiulpin et al. [28] trained a deep Siamese CNN using 18,376 plain radiographs and validated it on 5,960 images from another independent cohort. The images in both cohorts were graded according to the KL grading system. The model yielded an AUROC of 0.93 for OA

diagnosis ($KL \geq 2$). Developing standardized computer-based methods for detecting cartilage lesions at MRI would be beneficial to OA diagnosis because cartilage loss is relevant for OA. A fully automated deep CNN-based cartilage lesion detection system was developed by Liu et al. [29]. The result from two individual evaluations showed that the AI-based diagnosis tool has an overall diagnostic AUROC of 0.91 and good intraobserver agreement with a k statistic of 0.76.

Recently deep learning has been adopted to detect bone erosions. Bone erosion is a peri-inflammatory destructive bone lesion that is associated with RA severity and poor functional outcome. As RA progresses, the marginal erosions become visible on X-ray images. A deep CNN model was leveraged by Rohrbach et al. [30] to distinguish between eroded and not eroded joints. A limited accuracy (balanced accuracy of 47.5%) was achieved due to the high data imbalance. Besides the X-ray images, the high-resolution peripheral quantitative computer tomography (HRpQCT) is a promising tool to quantify the shape, volume, and surface area of erosions. Ren et al. [31] employed deep neural networks on HRpQCT scans to do segmentation of bone surface boundary and classification of the difference between healthy and eroded image patches. Patients with RA are at increased risk of developing osteoporosis. Kuo et al. investigated the potential of a deep texture encoding network (Deep-TEN) model for the low bone mineral density (BMD) estimation [32]. The Deep-TEN model was also used to assess the radiographic bone texture in the distal metacarpal bone [33] relevant to RA and knee OA [34].

Application in Mixture Data

Some studies in machine learning using integrated multi-type data for rheumatology research, instead of using a single type of data along. Orange et al. [35] employed k-means clustering and SVM on synovial histologic feature data and RNA sequencing data to identify three distinct molecular subtypes of RA synovial tissue correlated with specific clinical

phenotypes. To accurately predict treatment responses in RA patients, based on clinical profiles with additional genetic information, the Gaussian process regression model was shown to be the best performer that was correctly predicting 78% of 680 subjects' treatment response, with an AUROC of 0.66 [36]. In addition to multi-type data integration, multimodal machine learning is an alternative way to analyze the mixed data types. Tiulpin et al. [37] developed a multimodal machine learning-based OA progression prediction model. The model fused the predictions of KL grade, and disease progression from a CNN model and patient's characteristics, surgery history, and symptomatic assessment results by a gradient boosting machine classifier. The model yielded an AUROC of 0.79, which is better than a reference logistic regression (AUROC of 0.75).

Future Perspectives and Challenges

In general, the adoption of new technology such as AI in rheumatology lags behind the popularity of AI in other industries and subspecialties. Despite this, the introduction of AI to the field has produced much interesting research proving that AI can positively impact rheumatology. The application of AI focuses on the extraction and interpretation of EHRs, medical imaging, and handling of genetic and immunophenotyping data, and only selected disease categories have been exposed to AI/ML. In the coming years, AI will inevitably have a greater impact on rheumatology's research and clinical practice. However, unique challenges will emerge in the process of applying AI/ML models in rheumatology.

First, the diagnoses of rheumatic diseases are notoriously difficult owing to the need for multiple diagnostic modalities and dependence on physicians' or patients' subjective observations. For example, EULAR/ACR classification criteria for RA include the number of joint involvement and duration of symptoms, both of which are subject to detailed medical history-taking; however, the validation studies of the criteria found the

sensitivities from 0.50 to 0.60 and specificities from 0.88 to 0.97 [38]. This caveat needs to be considered for AI applications based on subjective data or those designed to assist the diagnosis of rheumatic diseases.

Second, many rheumatic diseases are relatively rare compared to common diseases. Systemic sclerosis, for example, has a prevalence ranging from 3.8 to 50 per 100,000 population [39]. Therefore, it is difficult to obtain enough high-quality data to support AI/ML models' training. In this regard, applying specific techniques to overcome the paucity of data is needed, such as fine-grained annotation, few-shot learning, or data augmentation.

Third, rheumatic diseases often have multi-organ involvement, so many patients need multidisciplinary care. For instance, patients with psoriatic arthritis need collaborative care by rheumatologists and dermatologists. The design of AI-assisted systems for these diseases need to consider the gaps in patient management between different subspecialties. The clinical heterogeneity may require researchers/developers to collect data from involved subspecialties, consult relevant experts, or update the model before applying it in a new clinical context.

Conclusions

The clinical application of AI/ML is still in infancy despite a growing number of high-quality research demonstrating the potential to support research and clinical work in rheumatology. The adoption of new technology needs to consider the general medical and subspecialty-specific caveats to maximize the benefits of AI/ML that may impact rheumatology care.

References

1. Lötsch J, Alfredsson L, Lampa J. Machine-learning-based knowledge discovery in rheumatoid arthritis-related registry data to identify predictors of persistent pain. *Pain*. 2020;161(1):114–26.
2. Chin C-Y, Hsieh S-Y, Tseng VS. eDRAM: effective early disease risk assessment with matrix factorization on a large-scale medical database: a case study on

- rheumatoid arthritis. *PLoS One.* 2018;13(11):e0207579.
3. Zhou S-M, Fernandez-Gutierrez F, Kennedy J, Cooksey R, Atkinson M, Denaxas S, et al. Defining disease phenotypes in primary care electronic health records by a machine learning approach: a case study in identifying rheumatoid arthritis. *PLoS One.* 2016;11(5):e0154515.
4. Murray SG, Avati A, Schmajuk G, Yazdany J. Automated and flexible identification of complex disease: building a model for systemic lupus erythematosus using noisy labeling. *J Am Med Inform Assoc.* 2019;26(1):61–5.
5. Lim J, Kim J, Cheon S. A deep neural network-based method for early detection of osteoarthritis using statistical data. *Int J Environ Res Public Health.* 2019;16(7):1281.
6. Lezcano-Valverde JM, Salazar F, León L, Toledano E, Jover JA, Fernandez-Gutierrez B, et al. Development and validation of a multivariate predictive model for rheumatoid arthritis mortality using a machine learning approach. *Sci Rep.* 2017;7(1):1–10.
7. Joo YB, Baek I-W, Park Y-J, Park K-S, Kim K-J. Machine learning-based prediction of radiographic progression in patients with axial spondyloarthritis. *Clin Rheumatol.* 2020;39(4):983–91.
8. Madrid-García A, Font-Urgelles J, Vega-Barbas M, León-Mateos L, Freites DD, Lajas CJ, et al. Outpatient readmission in rheumatology: a machine learning predictive model of patient's return to the clinic. *J Clin Med.* 2019;8(8):1156.
9. Norgeot B, Glicksberg BS, Trupin L, Lituiev D, Gianfrancesco M, Oskotsky B, et al. Assessment of a deep learning model based on electronic health record data to forecast clinical outcomes in patients with rheumatoid arthritis. *JAMA Netw Open.* 2019;2(3):e190606-e.
10. Reddy BK, Delen D. Predicting hospital readmission for lupus patients: an RNN-LSTM-based deep-learning methodology. *Comput Biol Med.* 2018;101:199–209.
11. Zheng C, Rashid N, Wu YL, Koblick R, Lin AT, Levy GD, et al. Using natural language processing and machine learning to identify gout flares from electronic clinical notes. *Arthritis Care Res.* 2014;66(11):1740–8.
12. Carroll RJ, Eyler AE, Denny JC, editors. Naïve electronic health record phenotype identification for rheumatoid arthritis. In: AMIA annual symposium proceedings. Washington DC, USA: American Medical Informatics Association; 2011.
13. Jorge A, Castro VM, Barnado A, Gainer V, Hong C, Cai T et al., editors. Identifying lupus patients in electronic health records: development and validation of machine learning algorithms and application of rule-based algorithms. Seminars in arthritis and rheumatism. Elsevier; 2019. <https://www.journals.elsevier.com/seminars-in-arthrits-and-rheumatism>
14. Chocholova E, Bertok T, Jane E, Lorencova L, Holazova A, Belicka L, et al. Glycomics meets artificial intelligence—potential of glycan analysis for identification of seropositive and seronegative rheumatoid arthritis patients revealed. *Clin Chim Acta.* 2018;481:49–55.
15. Ceccarelli F, Sciandrone M, Perricone C, Galvan G, Cipriano E, Galligari A, et al. Biomarkers of erosive arthritis in systemic lupus erythematosus: application of machine learning models. *PLoS One.* 2018;13(12):e0207926.
16. Poppenberg KE, Jiang K, Li L, Sun Y, Meng H, Wallace CA, et al. The feasibility of developing biomarkers from peripheral blood mononuclear cell RNAseq data in children with juvenile idiopathic arthritis using machine learning approaches. *Arthritis Res Ther.* 2019;21(1):1–10.
17. Patrick MT, Stuart PE, Raja K, Gudjonsson JE, Tejasvi T, Yang J, et al. Genetic signature to provide robust risk assessment of psoriatic arthritis development in psoriasis patients. *Nat Commun.* 2018;9(1):1–10.
18. Franks JM, Martyanov V, Cai G, Wang Y, Li Z, Wood TA, et al. A machine learning classifier for assigning individual patients with systemic sclerosis to intrinsic molecular subsets. *Arthritis Rheumatol.* 2019;71(10):1701–10.
19. Figgett WA, Monaghan K, Ng M, Alhamdoosh M, Maraskovsky E, Wilson NJ, et al. Machine learning applied to whole-blood RNA-sequencing data uncovers distinct subsets of patients with systemic lupus erythematosus. *Clin Transl Immunol.* 2019;8(12):e01093.
20. Robinson GA, Peng J, Dönnés P, Coelewij L, Naja M, Radziszewska A, et al. Disease-associated and patient-specific immune cell signatures in juvenile-onset systemic lupus erythematosus: patient stratification using a machine-learning approach. *Lancet Rheumatol.* 2020;2(8):e485–e96.
21. Plant D, Maciejewski M, Smith S, Nair N, Maximising Therapeutic Utility in Rheumatoid Arthritis Consortium tRSG, Hyrich K, et al. Profiling of gene expression biomarkers as a classifier of methotrexate nonresponse in patients with rheumatoid arthritis. *Arthritis Rheumatol.* 2019;71(5):678–84.
22. Kegerreis B, Catalina MD, Bachali P, Geraci NS, Labonte AC, Zeng C, et al. Machine learning approaches to predict lupus disease activity from gene expression data. *Sci Rep.* 2019;9(1):1–12.
23. Yang T, Zhu H, Gao X, Zhang Y, Hui Y, Wang F. Grading of metacarpophalangeal rheumatoid arthritis on ultrasound images using machine learning algorithms. *IEEE Access.* 2020;8:67137–46.
24. Tang J, Jin Z, Zhou X, Zhang W, Wu M, Shen Q, et al. Enhancing convolutional neural network scheme for rheumatoid arthritis grading with limited clinical data. *Chin Phys B.* 2019;28(3):038701.
25. Norman B, Pedroia V, Noworolski A, Link TM, Majumdar S. Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs. *J Digit Imaging.* 2019;32(3):471–7.

26. Andersen JKH, Pedersen JS, Laursen MS, Holtz K, Grauslund J, Savarimuthu TR, et al. Neural networks for automatic scoring of arthritis disease activity on ultrasound images. *RMD Open*. 2019;5(1):e000891.
27. Xue Y, Zhang R, Deng Y, Chen K, Jiang T. A preliminary examination of the diagnostic value of deep learning in hip osteoarthritis. *PLoS One*. 2017;12(6):e0178992.
28. Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. *Sci Rep*. 2018;8(1):1–10.
29. Liu F, Zhou Z, Samsonov A, Blankenbaker D, Larison W, Kanarek A, et al. Deep learning approach for evaluating knee MR images: achieving high diagnostic performance for cartilage lesion detection. *Radiology*. 2018;289(1):160–9.
30. Rohrbach J, Reinhard T, Sick B, Dürr O. Bone erosion scoring for rheumatoid arthritis with deep convolutional neural networks. *Comput Electr Eng*. 2019;78:472–81.
31. Ren J, Moaddel A, Hauge EM, Keller KK, Jensen RK, Lauze F, editors. Automatic detection and localization of bone erosion in hand HR-pQCT. Medical imaging 2019: computer-aided diagnosis. San Diego, California, USA: International Society for Optics and Photonics; 2019.
32. Kuo C, Miao S, Zheng K, Lu L, Hsieh C, Lin C, et al. OP0301 prediction of low bone mineral density and frax score by assessing hip bone texture with deep learning. London, UK: BMJ Publishing Group Ltd; 2020.
33. Kuo C, Miao S, Zheng K, Lu L, Hsieh C, Lin C. SAT0564 bone texture analysis with deep learning in hand radiographs for assessing the risk of rheumatoid arthritis. London, UK: BMJ Publishing Group Ltd; 2020.
34. Kuo C, Zheng K, Miao S, Lu L, Hsieh C, Lin C, et al. OP0062 predictive value of bone texture features extracted by deep learning models for the detection of osteoarthritis: data from the osteoarthritis initiative. London, UK: BMJ Publishing Group Ltd; 2020.
35. Orange DE, Agius P, DiCarlo EF, Robine N, Geiger H, Szymbonifka J, et al. Identification of three rheumatoid arthritis disease subtypes by machine learning integration of synovial histologic features and RNA sequencing data. *Arthritis Rheumatol*. 2018;70(5):690–701.
36. Guan Y, Zhang H, Quang D, Wang Z, Parker SC, Pappas DA, et al. Machine learning to predict anti-tumor necrosis factor drug responses of rheumatoid arthritis patients by integrating clinical and genetic markers. *Arthritis Rheumatol*. 2019;71(12):1987–96.
37. Tiulpin A, Klein S, Bierma-Zeinstra SM, Thevenot J, Rahtu E, van Meurs J, et al. Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. *Sci Rep*. 2019;9(1):1–11.
38. Bykerk VP, Massarotti EM. The new ACR/EULAR classification criteria for RA: how are the new criteria performing in the clinic? 2012;51(Suppl 6):vi10–5. <https://doi.org/10.1093/rheumatology/kes280>.
39. Zhong L, Pope M, Shen Y, Hernandez JJ, Wu L. Prevalence and incidence of systemic sclerosis: a systematic review and meta-analysis. *Int J Rheum Dis*. 2019;22(12):2096–107. <https://doi.org/10.1111/1756-185X.13716>.



AIM in Osteoporosis

56

Sokratis Makrogiannis and Keni Zheng

Contents

Introduction	786
Methods	786
Nonsparse Classification Techniques: Texture-Based, Patch-Based, and Deep Learning	786
Classifiers and Discriminant Functions	791
Bag of Keypoints	792
Deep Neural Networks	792
Sparse Representation and Classification	792
Integrative Ensemble Sparse Analysis Techniques	794
Results	797
Discussion	799
References	800

Abstract

In this chapter, we explore and evaluate methods for trabecular bone characterization and osteoporosis diagnosis with increased interest in sparse approximations. We first describe texture representation and classification techniques, patch-based methods such as Bag of Keypoints, and more recent deep neural

networks. Then we introduce the concept of sparse representations for pattern recognition and we detail integrative sparse analysis methods and classifier decision fusion methods. We report cross-validation results on osteoporosis datasets of bone radiographs and compare the results produced by the different categories of methods. We conclude that advances in the AI and machine learning fields have enabled the development of methods that can be used as diagnostic tools in clinical settings.

S. Makrogiannis (✉) · K. Zheng
Division of Physics, Engineering, Mathematics and
Computer Science, Delaware State University, Dover, DE,
USA
e-mail: smakrogiannis@desu.edu

Keywords

Sparse representation · Ensemble classifiers ·
Computer-aided diagnosis · Osteoporosis ·
Fracture risk

Introduction

Osteoporosis is a skeletal disorder characterized by decreased bone strength that may lead to susceptibility of fracture [1]. There are more than 3 millions of people diagnosed with osteoporosis in the USA per year. The risk is increasing with age, especially the people who are over 40. Timely diagnosis of osteoporosis can effectively predict fracture risk and allow for effective treatment.

Trabecular bone characterization and automated and accurate diagnosis of bone osteoporosis is significant for improving public health. Aerial Bone Mineral Density (BMD) is computed in dual-energy X-ray absorptiometry (DXA) scans to diagnose osteoporosis [2]. However, BMD can predict fracture with only 60% accuracy. Analysis of trabecular bone micro-architecture can significantly improve the prediction rates, but this information requires bone biopsy with histomorphometric analysis. The task of obtaining trabecular bone micro-architecture information by noninvasive methods is a nontrivial scientific problem [3]. Diagnosis of osteoporosis using bone radiograph scans presents some challenges, mainly because images of osteoporotic and healthy subjects are visually very similar. Previous approaches to evaluating bone structure on radiographs by 2D texture analysis were reported in [2, 4, 5]. Moreover, in [6, 7] the authors propose to use 2D texture analysis to characterize 3D bone microarchitecture.

Here we present and evaluate mathematical methods and algorithms for computer-aided diagnosis. The application domain is osteoporosis diagnosis in radiographs of the calcaneus bone. We will explore the use of sparse modeling and classification for classifying diseased from healthy subjects. Then we will present ensemble sparse techniques to find more accurate solutions than individual classification techniques. We will also test other classification techniques based on texture features, or patch-based techniques such as the Bag of Keypoints and deep learning methods.

Methods

Nonsparse Classification Techniques: Texture-Based, Patch-Based, and Deep Learning

Introduction to Texture-Based Classification

Texture is an image property that can be used for segmenting and classifying images into different objects. We can define texture as a structure consisting of a group of related elements [8]. The pixels in this group are called texture primitives, texture elements, or texels.

Texture analysis techniques are mainly applied to texture recognition and texture-based shape analysis [8]. Generally, people consider texture as fine when the texture element is small and there are large differences between elements, and coarse when the element is large and there are only few elements in the image, grained and smooth. For scientific applications of texture, we use more precise characteristics such as tone and structure [9]. Tone is more about pixel intensity and structure is about the spatial relationship between texture elements. There are many methods for texture extraction, such as wavelet analysis, Gabor filters, co-occurrence matrices, intensity histogram-based, and spatial frequency domain descriptors.

We present texture-based methods for computer-aided diagnosis of diseased and healthy subjects [10]. Our premise is that the deterioration of disease can be captured by textural features. We first computed texture features based on wavelet decomposition, discrete Fourier and Cosine transforms, fractal dimension, statistical co-occurrence indices, and structural texture descriptors. We employed feature selection techniques that consider the individual feature predictive ability and inter-feature redundancy to find the most discriminant feature set. In the classification stage, we employed Naïve Bayes, Multilayer Perceptron, Bayes Network, Random Forests, and Bagging models for diagnosis.

Feature Computation The purpose of this stage is to compute texture descriptors that can be used for separation between groups of healthy and diseased subjects. This is usually performed

in a high-dimensional feature space to reduce the Bayes error rate. Next, we describe frequently used feature sets.

Fractal Dimension These features have shown promise in texture classification applications. A fractal is defined as a mathematical set whose Hausdorff dimension exceeds the fractal's topological dimension [11]. It has been shown that fractal dimension correlates well with a function's roughness. Therefore, we used fractal dimension to measure the roughness and granularity of the image intensity function. The topological dimension of this function is equal to 3, consisting of 2 spatial dimensions plus the intensity.

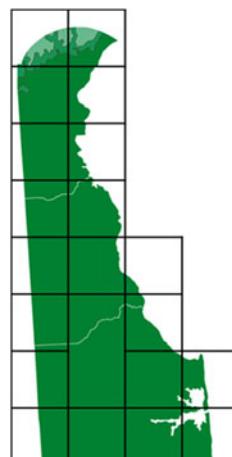
The method of box counting can be utilized to compute the fractal dimension. Assuming a fractal structure with dimension D , we let $N(\epsilon)$ be the number of nonempty boxes of size ϵ required to cover the fractal support. Using the relation $N(\epsilon) \simeq \epsilon^{(-D)}$, we can numerically estimate D from

$$D = \lim_{\epsilon \rightarrow 0} \frac{\log N(\epsilon)}{-\log \epsilon} \quad (1)$$

by least squares fitting. We display an example of application of box counting in Fig. 1.

For the case of grayscale images or continuous functions, we generated eight binary sets using multiple Otsu thresholding, then computed the fractal dimension, area, and mean intensity for each point set as in [12].

Fig. 1 Box counting to compute the fractal dimension of Delaware state boundary



Wavelet Texture Descriptors A multiscale texture descriptor is usually very useful for classification. Gabor filter banks and wavelet transforms are both multiscale spatial-spatial frequency filtering techniques. The discrete wavelet transform is frequently applied using tree or pyramid hierarchies for texture representation. Multiband analysis offers advantages over the traditional discrete Fourier transform, but wavelet transform does not produce as exact a result as the Fourier transform.

Discrete Wavelet Frames Discrete wavelet frames employ a filter bank for multiscale decomposition. The Haar wavelet with a low-pass filter

$$H(z) = (1 + z^{-1})/2 \quad (2)$$

and a corresponding high-pass filter

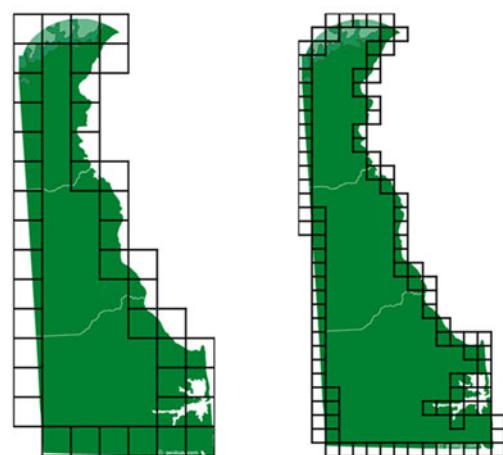
$$G(z) = (1 - z^{-1})/2 \quad (3)$$

is frequently used because of its efficiency and computational simplicity.

The largest filter kernels will have size $2^{maxlevel}$, where the $maxlevel$ is the number of multiresolution levels. At each level, we filter the image by using the filter combinations:

$$H_x H_y, H_x G_y, G_x H_y, G_x G_y, \quad (4)$$

where H_x is the low-pass filter along the x direction and G_y is the high-pass filter along the y direction.



To produce the wavelet frame representation, we compute the discrete wavelet transform for all possible signal shifts at multiple scales. The filters are used to decompose the image into subbands. We compute the orthogonal projections and residuals for a full discrete wavelet expansion. We then compute energy, variance, entropy, contrast, skewness, and kurtosis signatures to form the texture descriptor. These characteristics are calculated as follows.

Contrast It measures the intensity contrast between a pixel $p(i, j)$ and its neighbors in an image by

$$\sum_{i,j} |i - j|^2 p(i, j) \quad (5)$$

Energy It is expressed by the sum of squared elements

$$\sum_{i,j} p(i, j)^2 \quad (6)$$

Skewness It measures the lack of symmetry. For a random variable x , the skewness is the third standardized moment γ_1

$$\gamma_1 = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{E[(X-\mu)^3]}{\left(E[(X-\mu)^2]\right)^{3/2}} = \frac{k_3}{k_2^{3/2}} \quad (7)$$

where μ is mean, σ is standard deviation, μ_3 is central moment, E is expectation operator, and k_i is the i th cumulants.

Kurtosis It measures the degree to which data points follow a heavy-tailed or light-tailed distribution. Higher kurtosis values correspond to heavier-tailed distributions.

$$Kurt[X] = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4} = \frac{E[(X-\mu)^4]}{\left(E[(X-\mu)^2]\right)^2} \quad (8)$$

Entropy It presents the state of a system, such as the disorder and randomness of the system. The wavelet entropy is defined in [13] as

$$S(p) = -\sum_j p_j \ln p_j \quad (9)$$

Wavelet Gabor Filter Bank The Gabor filter is a linear filter that can extract relevant characteristics for multiple frequencies and orientations (Fig. 2), similarly to the human visual system.

Gabor functions form a complete but non-orthogonal basis. In the spatial domain, a 2D Gabor filter is a Gaussian kernel function modulated by a sinusoidal plane wave. Gabor filters are often used for texture identification, and good results have been achieved. The filter is represented in complex form as follows:

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{w^2 + \gamma^2 v^2}{2\sigma^2}\right) \exp\left(i\left(2\pi\frac{w}{\lambda} + \psi\right)\right) \quad (10)$$

and

$$w = x \cos \theta + y \sin \theta \quad (11)$$

$$v = -x \sin \theta + y \cos \theta \quad (12)$$

where λ is the wavelength of the sinusoidal factor, θ is the orientation of the normal to the parallel stripes of a Gabor function, ψ is the phase offset, σ is the standard deviation of the Gaussian envelope, and γ is the spatial aspect ratio.

The filter dictionary can be produced by dilations and rotations of the mother Gabor wavelet.

Local Binary Patterns (LBP) For each pixel pix in the image, we compare the intensity of pix to the intensities of its eight neighbors. If the intensity of pix is greater or equal to its i th (where $i = 1, 2, \dots, 8$) neighbor, we set $b_i = 0$, otherwise $b_i = 1$. From these eight neighbors, we construct an eight-digit binary number $b_1 b_2 b_3 b_4 b_5 b_6 b_7 b_8$. We use the histogram of these numbers as a texture descriptor [14].

Discrete Fourier and Cosine Transforms The Discrete Fourier transform and the Discrete Cosine transform coefficients aim to capture characteristics of texture in the spatial frequency domain. For example, fine texture has greater high frequency components, whereas coarse texture is represented

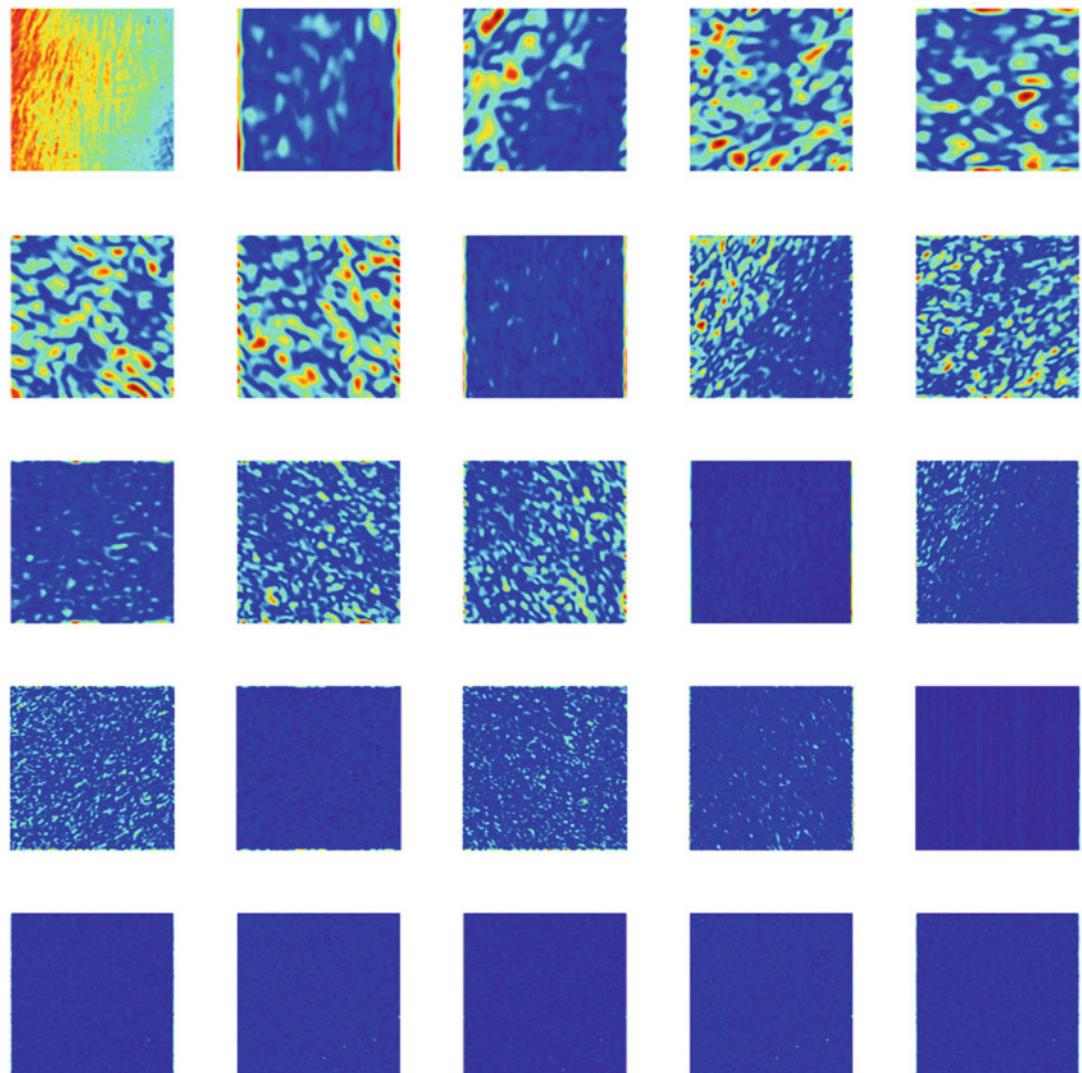


Fig. 2 The original bone radiograph and the Gabor texture components of a healthy subject using four scales and six orientations. The 24 components are calculated on the original image in top left. While these maps pronounce

by lower frequencies. The Discrete Fourier and Cosine transforms are defined as follows,

Discrete Fourier transform (DFT):

$$F(k, l) = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f(m, n) \cdot e^{-j2\pi(\frac{mk}{M} + \frac{nl}{N})}, \quad (13)$$

where $f(m, n)$ is the pixel intensity at (m, n) , and $k = 0, 1, 2, \dots, N-1, l = 0, 1, 2, \dots, M-1$.

the texture characteristics, visual interpretation is still particularly challenging. Therefore, a machine learning technique is needed to distinguish healthy from osteoporotic subjects

Discrete Cosine Transform (DCT) uses only cosine basis functions:

$$C(k, l) = \sqrt{\frac{\alpha}{MN}} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f(m, n) \cdot \cos \frac{\pi(2m+1)k}{2M} \cos \frac{\pi(2n+1)l}{2N}, \quad (14)$$

where $\alpha = 1$, if $k = l = 0$; $\alpha = 4$, if $1 \leq k \leq M-1, 1 \leq l \leq N-1$.

We use the 8×8 coefficients corresponding to lower frequencies for classification.

Law's Texture Energy Masks The texture energy is computed by a set of 5×5 convolution masks (level, edges, waves, spots, and ripples) to measure the amount of variation within a fixed-size window. We use the average level (intensity) feature to normalize intensity range and then we use the remaining 24 components to form the texture vector. Next, we calculate the mean, variance, energy, skewness, kurtosis, and entropy for each component.

Edge Histogram We compute the intensity gradient magnitude $|\nabla f|$ and then calculate its histogram by

$$p_{|\nabla f|}(|\nabla f| = r_k) = \frac{n_k}{N}, k = 0, \dots, L - 1 \quad (15)$$

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_N} \right)^T \quad (16)$$

Gray Level Co-Occurrence Matrix (GLCM)

The GLCM calculates how the frequency of occurrence of gray-level pairs (i, j) in horizontal, vertical, or diagonal pixel adjacencies on the image plane, displayed in Fig. 3. Horizontal (0°), vertical (90°), and diagonal (-45° , -135°) dimensions of analysis are denoted by P_0 , P_{90} , P_{45} , and P_{135} , respectively. After we create the GLCMs, we compute contrast, correlation, energy, and homogeneity measures.

Feature Selection This classification component aims to select relevant and informative

features for classification. It is applied to improve classification performance, to reduce computational complexity, and to interpret data.

Correlation-based Feature Selection (CFS)

This method selects features that are highly correlated with the pattern classes, but have low correlation with the remaining features. The subset evaluation function is given by:

$$Merit_S = \frac{\bar{k}_{rf}}{\sqrt{k + k(k-1)r_{ff}}} \quad (17)$$

where $Merit_S$ is the merit of the selected feature set S , \bar{k}_{rf} is the mean correlation between the features and class with $f \in S$, and r_{ff} is the mean pairwise feature correlation. The numerator expresses predictive capacity, while the denominator expresses feature redundancy.

Best First Search (BF) Searches the space of feature subsets by greedy hillclimbing that may include backtracking. Best first may search forward, or backward, or, consider all possible single feature additions and deletions at a given point using a bi-directional strategy.

Genetic Algorithm-based Search (GA) Genetic search works by having a population of variables representing feature sets and performs the operations of reproduction, cross-over, and mutation in each generation to get the offspring that optimizes a feature set-related objective function.

Information Gain (IG) This function measures the information gain with respect to the class:

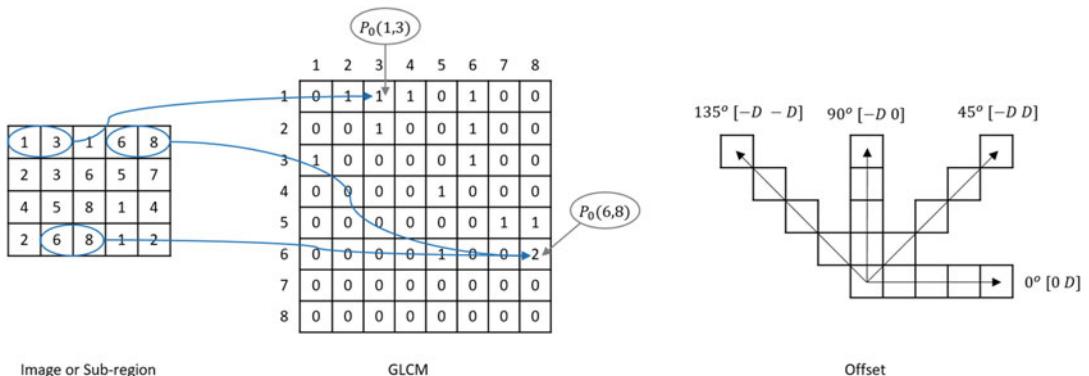


Fig. 3 Process used to create the GLCM (left) and Offset of GLCM (right)

$$\begin{aligned} \text{InfoGain(Class, Attribute)} \\ = H(\text{Class}) - H(\text{Class}|\text{Attribute}) \end{aligned} \quad (18)$$

where H is the entropy of each class given by $H(\text{Class}) = -p_{\text{Class}} \log p_{\text{Class}}$. We select the attributes by individual ranking evaluation.

Ranker Using Ranker as a search means that we will rank the features based on the features' individual evaluations. A threshold can be set in Ranker, and features that are smaller than this threshold will be removed from the feature set. Ranker is used with attribute evaluators, such as Information Gain (IG), feature selection and entropy, etc.

Classifiers and Discriminant Functions

Naïve Bayes (NB) This model assumes conditional statistical independence $p(x|\omega_j) = \prod_{k=1}^D p(x_k|\omega_j)$ where $x = (x_1, x_2, \dots, x_D)^T$ and D is the dimensionality of the feature space. The posterior probability is based on Bayes' formula:

$$p(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)}.$$

The MAP decision rule is typically used for classification. Suppose we have two categories ω_1 and ω_2 with discriminant functions $g_1(x)$, $g_2(x)$, where

$$\begin{aligned} g_i(x) = -\frac{1}{2}(x - \mu_i)^T \sum_i^{-1} (x - \mu_i) + \frac{D}{2} \ln 2\pi \\ - \frac{1}{2} \ln \left| \sum_i \right| + \ln P(\omega_i). \end{aligned}$$

Then we can define a single discriminant function by

$$g(x) = g_1(x) - g_2(x).$$

The decision rule is:

$$\begin{cases} \omega_1, & \text{if } g(x) > 0 \\ \omega_2, & \text{if } g(x) < 0 \end{cases}.$$

Multilayer Perceptron (MLP) A multilayer perceptron is a feedforward artificial neural

network system that maps input patterns onto class labels. An MLP has multiple layers of nodes that are fully connected to the next layer. Each node is a neuron with a nonlinear activation function. MLP utilizes backpropagation for supervised learning [15, 16]. Because MLP has multiple layers of logistic regression models, it can distinguish data that are not linearly separable. In learning by backpropagation that can be considered as an extension of the LMS algorithm, we adjust the connection weights, according to the amount of error in the output compared to the expected result.

Bayes Network (BN) A Bayes network is a probabilistic graphical model that uses a directed acyclic graph to represent a set of random variables and their conditional dependencies. In a Bayesian network, the joint probability density function can be written as the product of univariate conditional density functions dependent on their parent variables:

$$p(x) = \prod_{v \in V_p} (x_v | x_{pa(v)}) \quad (19)$$

where $pa(v)$ denotes the parents of v . In the graph, the parents are vertices directly connected to v by a single edge.

Bagging For a training set S with size k , Bagging generates j training subsets denoted as S_i with size $k' < k$, by sampling from S uniformly and with replacement. We denote the original set as A . In the training stage, we first have $D = \emptyset$ and j is the number of classifiers to train. Then for $p = 1, 2, \dots, j$, we take a bootstrap sample S_p from A to train classifier D_p . Then the classifier D_p is added to the current ensemble, $D = D \cup D_p$. The class label prediction for the input x is obtained by majority voting on the individual classifier decisions produced by D_1, \dots, D_j [16].

Random Forests (RF) Random forests is an ensemble learning method that constructs multiple decision trees from subsets of the training set and uses random feature selection for node splitting. RF decides the class after applying voting to the predicted classes by the individual trees for classification, or by calculating the mean prediction for regression. Random forests address

the overfitting tendency of the decision trees and have shown robustness with respect to noise [17].

Bag of Keypoints

Bag of Keypoints (BoK) [18] is a patch-based technique that originates from Bag of Features methods. These methods have been applied to image recognition and classification and have produced very good results. They apply feature detection, extraction, and clustering for finding the most representative features in the training database. In the next step, they build a vocabulary that consists of the frequency of occurrence of these features. In the testing stage, features are extracted from the unlabeled image and encoded using the vocabulary that was built during training. Then a learning method is applied to classify the test pattern into one of the classes.

We employed the support vector machine (SVM) classifier for learning a discriminant function from the encoded features and classifying unlabeled samples. In the SVM module, we evaluated the use of linear or radial basis function kernels. We utilized radial basis function kernels for our experiments to address possible non-linearity of the decision boundary.

Deep Neural Networks

Deep learning methods and more specifically convolutional neural networks have recently re-emerged as powerful techniques for image segmentation, object recognition, and classification [19–23]. These techniques simultaneously learn the set of features and the decision function. In contrast to traditional texture-based techniques, deep networks do not need to receive a hand-crafted feature set as input. Deep learning methods have been applied to biomedical image datasets and have produced very good results.

We employed sequential and residual networks of varying complexity such as Alexnet [19], Googlenet [20], Resnet18 [23], and Inceptionv3

[21]. Because our datasets are small, we employed transfer learning techniques to adjust the weights of pretrained networks, instead of learning the decision function from the beginning as described in [22, 24]. All networks were pretrained on Imagenet that is a database of 1.2 million natural images.

We applied transfer learning to each network in slightly different ways. To adjust Alexnet to our data, we replaced the pretrained fully connected layers with three new fully connected layers. We set the learning rates of the pretrained layers to 0 to keep the network weights fixed, and we trained the new fully connected layers only. In the case of Googlenet, we set the learning rates of the bottom 10 layers to 0, we replaced the top fully connected layer with a new fully connected layer, and we assigned a greater learning rate factor for the new layer than the pretrained layers.

To provide the networks with additional training examples, we employed data resampling using randomly centered patches, followed by data augmentation by rotation, scaling, horizontal flipping, and vertical flipping. Finally, we applied hyperparameter tuning using Bayesian optimization to find the optimal learning rate, mini-batch size and number of epochs.

Sparse Representation and Classification

The concept of sparsity has been used in many methods of mathematics, computer science, and engineering and plays an important role in machine learning and pattern recognition. Next, we introduce the standard sparse technique, the details of this method, and other related sparse techniques.

Overview of Sparse Modeling Methods Tissue classification is typically achieved by supervised machine learning approaches. Among numerous techniques that proposed generative or discriminative models, use of kernels, and linear or nonlinear approaches, sparse classification techniques have shown promise and applicability for characterizing visual patterns in region of

interest (ROI)-based analyses. Sparse representation techniques have been applied to extensive fields including coding, feature extraction and classification, superresolution [25], and regularization of inverse problems [26]. Exploration of signal's sparsity may provide insight into the important patterns of prototyping of objects category. The sparse representation is more concise for compression and naturally discriminative for classification [27]. Sparse representation techniques calculate a sparse linear combination of atoms for describing a vector sample using an overcomplete dictionary of prototypes. If the representations of these linear combinations are sufficiently sparse, then they can be used for object recognition and classification of imaging patterns.

The authors in [27] proposed the sparse representation classification (SRC) method to recognize 2,414 frontal-face images of 38 individuals of Yale B Database and over 4,000 frontal images for 126 individuals of AR Database, producing recognition rates greater than 90% for both databases. Another notable face recognition application of sparse coding was published in [28] reporting high levels of classification accuracy. Dictionary learning techniques have also emerged as solutions for sparse representation in the recent years. The utilization of K-SVD, where SVD denotes singular-value decomposition, for dictionary learning has been studied to produce a dictionary aiming for more accurate representation [29]. In [30], the K-SVD technique has been used for color image restoration to handle non-homogeneous noise and information missing problems. The authors in [31] observed that K-means may yield as good precision rate as K-SVD when we use the same number of atoms. The SRC method with dictionary learning was applied to classification of pulmonary patterns of diffuse lung disease in [32]. Additional algorithms such as matching pursuit (MP), orthogonal matching pursuit (OMP), and basis pursuit (BP) have been proposed for codebook design [29].

Sparse Representation and Classification

Sparse representation or approximation techniques construct a dictionary from labeled training samples to calculate a linear representation of a

test sample. This representation can be used to make a decision for the class of the test sample. Assuming that a dataset has k distinct classes, s samples, and for i th class, there are s_i samples, so that $s = \sum_i s_i$, we define a dictionary matrix M from the training set as

$$M = [v_{1,1}, v_{1,2}, \dots, v_{k,s_k}]. \quad (20)$$

where $M \in \mathbb{R}^{l \times s}$, and $v_{i,h}$ is a column vector for the h th sample from i th class. In image classification applications, a $p \times q$ grayscale image forms a vector $v \in \mathbb{R}^l$, $l = p \times q$ using lexicographical ordering.

A new test sample $y \in \mathbb{R}^l$, can be represented by a linear combination of samples $y = \sum_{i=1}^k \beta_{i,1} v_{i,1} + \beta_{i,2} v_{i,2} + \dots + \beta_{i,s_i} v_{i,s_i}$, where $\beta_{i,h} \in \mathbb{R}$ are scalar coefficients. Hence, the test sample y can be rewritten as:

$$y = Mx_0 \in \mathbb{R}^l. \quad (21)$$

where x_0 is a sparse solution. If there are sufficient training samples, the components of x_0 are equal to zero except for the components corresponding to the i th class. Then $x_0 = [0, 0, \dots, \beta_{i,1}, \beta_{i,2}, \dots, \beta_{i,s_i}, 0, 0, \dots, 0]^T \in \mathbb{R}^s$.

In [33], it was proved that whenever $y = Mx$ for some x , if there are less than $l/2$ nonzero entries in x , x is the unique sparse solution: $\hat{x}_0 = x$. Finding an accurate sparse representation of an underdetermined system of linear equations is an NP-hard problem [34]; therefore, only approximate solutions can be found. The authors in [35] supported that if the solution x_0 is sparse enough, it is equal to the solution \hat{x}_1 of the l^1 -minimization problem:

$$(l^1) : \hat{x}_1 = \arg \min \|x\|_1 \text{s.t. } Mx = y. \quad (22)$$

In sparse representation classification, we define a characteristic function $\delta_i : \mathbb{R}^s \rightarrow \mathbb{R}^s$ that has nonzero entries, only if x is associated with class i . Then the function $\hat{y}_i = M\delta_i(\hat{x}_1)$ represents the given sample y using components from class i only. To classify y and determine the class label $\hat{\omega}_i$, we minimize the residual between y and \hat{y}_i [27]:

$$\hat{\omega}_i = \arg \min_i r_i(y) \doteq \|y - M\delta_i(\hat{x}_1)\|_2. \quad (23)$$

This technique also adopts the sparsity concentration index (SCI) to measure the efficiency of class-conditional representation of a sample. The SCI of a coefficient vector $x \in \mathbb{R}^s$ is $SCI(x) = \frac{k \times \max_i \|\delta_i(x)\|_1 / \|x\|_1 - 1}{k-1} \in [0, 1]$ as defined in [27]. For a solution \hat{x} , if $SCI(\hat{x})$ is 1, y is only represented by images from a single class, and if $SCI(\hat{x}) = 0$, the components of β are spread evenly over all classes.

Algorithms for Solving the Sparse Representation Problem Earlier in this section, we mentioned that finding an accurate solution of sparse representation is an NP hard problem and described a method for finding an approximate solution to Eq. (22). Here we outline three common methods for solving (22). One is the matching pursuit (MP) method. MP selects atoms, one at a time, to minimize the approximation error. This is done by finding the atom with the largest inner product of the signal, subtracting the approximation from the signal using only that atom, repeat this step until it finds the satisfying residual. In [36] the authors proposed an algorithm as orthogonal matching pursuit (OMP). OMP modified MP to achieve full backward orthogonality of residuals (errors) at each step, resulting in improved convergence. Another optimization method for this problem is basis pursuit (BP) [37]. This method optimizes a joint expression of the constraint and the objective function and is equivalent to the LASSO method.

Second Order Cone Programming Formulation The second order cone (SOCP) programming problems are convex optimization problems. The SOCP can be used to implement linear programming (LP), convex quadratic programs (QPs), and convex quadratically constrained quadratic programs (QCQPs) [38]. The standard form of SOCP is defined as following:

$$\min u_1^\top x_1 + \cdots + u_n^\top x_n \quad (24)$$

$$s.t. A_1 x_1 + \cdots + A_n x_n = b \quad (25)$$

$$x_i \geq 0 \text{ for } i = 1, 2, \dots, n \quad (26)$$

SOCP can be used for solving problems of the form,

$$\min_x f(x) \text{ s.t. } \begin{cases} c(x) \leq 0 \\ ceq(x) = 0 \\ A \cdot x \leq b \\ Aeq \cdot x = beq \\ lb \leq x \leq ub \end{cases} \quad (27)$$

where $f(x)$ is an objective function, lb and ub are lower bound and upper bound respectively, A is a matrix and b is a vector for inequality, Aeq is a matrix and beq is a vector for equality, and $c(x)$ and $ceq(x)$ are constraint functions that return vectors. Especially, $f(x)$, $c(x)$, and $ceq(x)$ can be nonlinear functions.

Integrative Ensemble Sparse Analysis Techniques

This method builds an ensemble of sparse representation classifiers based on block decomposition of the input ROI to address shortcomings caused by high dimensionality and to introduce spatial localization in sparse approximations [39]. Figure 4 summarizes the main stages of our method that may be divided in block-based learning and Bayesian model averaging to form decision functions.

Block Decomposition We first divide each training ROI into non overlapping blocks of size $m \times n$. Thus, each ROI image is expressed as $I = [B^1, B^2, \dots, B^{NB}]$, where NB is the number of blocks in an image. The dictionary D^j , where $j = 1, 2, \dots, NB$ corresponds to the block B^j at the same index within the image ROI. The dictionary D^j for all the s images can be represented as follows:

$$D^j = [bv_{1,1}^j, bv_{1,2}^j, \dots, bv_{k,s_k}^j], \quad (28)$$

where $bv_{i,h}^j$ is the column vector denoting the h th sample, i th class, j th block B^j .

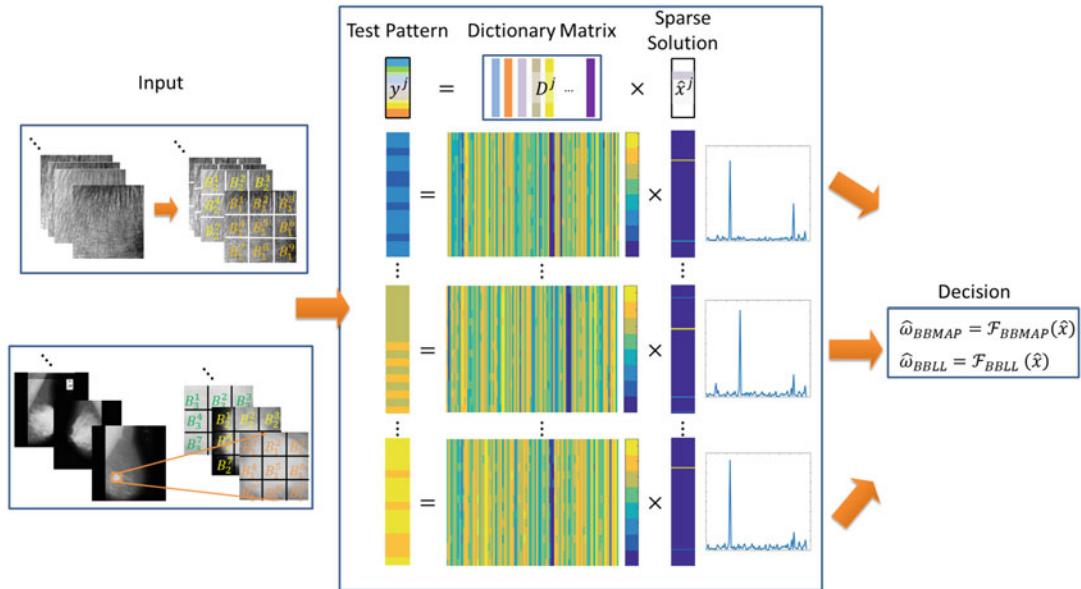


Fig. 4 Main stages of our integrative sparse modeling system: block-based analysis, sparse solutions, and decision functions

Ensemble Classification In this stage, each test sample is classified by constructing ensembles of classifiers that solve a set of sparse coding and classification problems, or hypotheses corresponding to the block components. Given a test sample y^j in j th block, we find the solution x^j of the regularized noisy ℓ^1 -minimization problem:

$$\hat{x}^j = \arg \min \|x^j\|_1 \text{ subject to } \|D^j x - y^j\|_2 \leq \epsilon \quad (29)$$

where $j = 1, 2, \dots, NB$. The test sample y^j will be assigned to the class ω_i^j , which has minimum approximation error calculated by (30).

$$\omega_i = \arg \min_i r_i(\hat{x}) \doteq \arg \min_i \|y - \hat{y}_i\|_2. \quad (30)$$

We utilize ensemble learning techniques in a Bayesian probabilistic setting as weighted sums of classifier predictions. We employ a function that applies majority voting to individual hypotheses (BBMAP) and an ensemble of log likelihood scores computed from relative sparsity scores (BBLL).

Maximum a Posteriori decision function (BBMAP) The class label for each test sample is determined by voting over the ensemble of $N B$ block-based classifiers. The predicted class label $\hat{\omega}$ is given by

$$\hat{\omega}_{BBMAP} = \mathcal{F}_{BBMAP}(\hat{x}) \doteq \arg \max_i pr(\omega_i | \hat{x}), \quad (31)$$

where \hat{x} is the composite extracted feature from the test sample given by the solution of (29). The probability for classifying \hat{x} into class ω_i is

$$pr(\omega_i | \hat{x}) = \sum_j^{NB} ND_{\omega_i^j} / NB \quad (32)$$

$$ND_{\omega_i^j} = \begin{cases} 1, & \text{if } \hat{x}^j \in i\text{th class} \\ 0, & \text{otherwise} \end{cases}, \quad (33)$$

where $ND_{\omega_i^j}$ is an indicator function whose values are determined by the individual classifier decisions.

Log likelihood approximation residual-based decision function (BBLL-R) We define a

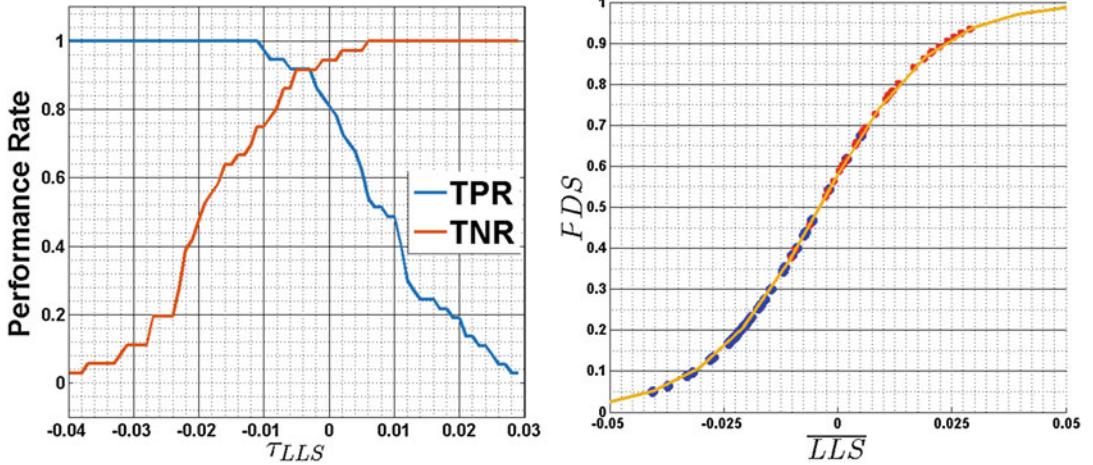


Fig. 5 An example of TPR and TNR curves versus τ_{LLS} for determining $\tau_{LLS}^* = c$ (left) and the sigmoid probability decision score PDS after calculating the parameters m, c for (38) (right)

likelihood score based on the residuals r_m ; r_n calculated in the sparse representation stage of each classifier. We calculate the expectation of $\overline{LLS(\hat{x})}$ over all individual classifiers:

$$\begin{aligned} \overline{LLS(\hat{x})} &= -\frac{1}{NB} \\ &\times \left[\sum_j^{NB} \log r_m^j(\hat{x}) - \sum_j^{NB} \log r_n^j(\hat{x}) \right], \end{aligned} \quad (34)$$

where $r_\omega^j(y)$ is the approximation residual for class ω and j is the block index:

$$r_\omega^j(y) = \|y - M\delta_\omega(\hat{x}_1)\|_2 \text{ for } j = 1, \dots, k. \quad (35)$$

Log likelihood sparsity-based decision function (BBLL-S) We define a likelihood score based on the relative sparsity scores $\|\delta_m(\hat{x}^j)\|_1$, $\|\delta_n(\hat{x}^j)\|_1$ calculated in the sparse representation stage of each classifier. We calculate the expectation of $LLS(x)$ over all individual classifiers:

$$\begin{aligned} \overline{LLS(\hat{x})} &= -\frac{1}{NB} \left[\sum_j^{NB} \log \|\delta_m(\hat{x}^j)\|_1 \right. \\ &\quad \left. - \sum_j^{NB} \log \|\delta_n(\hat{x}^j)\|_1 \right], \end{aligned} \quad (36)$$

The introduction of the log-likelihood score accommodates the definition of a decision function for the state $\hat{\omega}$. To determine the class, we apply a decision threshold τ_{LLS} to $\overline{LLS(\hat{x})}$.

$$\begin{aligned} \hat{\omega}_{BBLL} &= \mathcal{F}_{BBLL}(\hat{x}) \\ &\stackrel{.}{=} \begin{cases} m\text{th class}, & \text{if } \overline{LLS(\hat{x})} \geq \tau_{LLS} \\ n\text{th class}, & \text{otherwise} \end{cases}. \end{aligned} \quad (37)$$

This threshold is expected to be equal to 0, if there is no estimation bias, but may be experimentally determined as the minimizer of a Bayes-type risk function. Hence, the optimal τ_{LLS}^* value can be determined by sampling the domain of τ_{LLS} and calculating true positive and true negative rates. Next, the optimal value is determined by the intersection of TPR and TNR curves. An example of this procedure for determining τ_{LLS}^* is displayed in Fig. 5.

In the next stage, we aim to convert the log likelihood decision scores to bounded posterior probability values using a sigmoid function. This function is denoted by Probability Decision Score (PDS) and is expressed by

$$PDS(\overline{LLS}) = \frac{1}{1 + \exp[-m(\overline{LLS} - c)]} \quad (38)$$

To calculate the model parameter c , we require that this function be equal to 50% probability for τ_{LLS}^* , hence $c = \tau_{LLS}^*$. To estimate m , we set a fixed probability level PDS_{min} (e.g., 5%, 10%) for the smallest value \overline{LLS}_{min} .

$$m = \frac{1}{\tau_{LLS}^* - \overline{LLS}_{min}} \ln \left(\frac{100 - PDS_{min}}{PDS_{min}} \right) \quad (39)$$

In Fig. 5, we display the graph of PDS versus LLS for one experiment. We can use PDS to express margins of uncertainty for classification in percentiles.

Results

The main goal of the presented experiments is to test the hypothesis that ensembles of block-based sparse classifiers improve the classification performance of conventional sparse representation. The second goal is to compare the proposed technique to classifiers based on texture, Bag of Keypoints, and deep learning. Finally, we compare the performances of the two decision functions BBMAP and BBLL. We describe the application of our system to osteoporosis diagnosis, report the classification results produced by leave-one-out cross-validation experiments, and discuss our findings.

Data Description The objective is to distinguish between healthy and osteoporotic subjects. The TCB challenge dataset contains labeled digital radiographs of 87 healthy and 87 osteoporotic subjects for training and testing (available online in <http://www.univ-orleans.fr/i3mto/data>, last access in 05/2018). The calcaneus trabecular bone images in the dataset have an ROI size of 400×400 pixels. A more detailed description of the dataset is provided in [40]. The experimental procedures involving human subjects were approved by the Institutional Review Board of the institution that provides the data.

Texture-Based Classification In the performance evaluation of conventional texture-based techniques, we calculated 723 texture-related features [10]. We selected features using correlation-based feature selection with best first search (CFS-BF), correlation-based feature selection with genetic algorithm

search (CFS-GA), information gain (IG), and no-feature selection as described in section “[Non-sparse Classification Techniques: Texture-Based, Patch-Based, and Deep Learning](#).” CFS-GA yielded an overall better performance than CFS-BF, IG, and no-feature selection on leave-one-out cross-validation (Table 1). This implies that CFS-GA effectively selects distinguishing features from the entire set. Among the tested classifiers, Bagging accomplished the highest performance with an ACC of 67.8% on leave-one-out cross-validation.

Bag of Keypoints Classifier The main parameters that we tuned were the fraction of features to keep for building the vocabulary, the vocabulary size, the penalty coefficient for misclassification of training samples in SVM, and the kernel scale. The results showed that BoK was able to separate successfully healthy from osteoporotic subjects with an ACC of 99.3% leave-one-out cross-validation as displayed in Table 2. This very high accuracy may be attributed to the extraction of discriminant features from the textured areas. Also, the employed SVM model is known to address data complexity caused by nonlinearity and high dimensionality.

Deep Neural Networks We evaluated the performance of the networks described in section “[Bag of Keypoints](#).” We set the learning rates of the convolutional layers to much lower values than the final layers. In this way, we largely preserved the pretrained layer weights at the initial and intermediate convolutional stages. The main parameters that we tuned were the learning rate, learning rate drop, size of mini-batch, and the number of epochs. We utilized grid search and Bayesian optimization search for parameter tuning. Among the deep neural networks, Resnet18 yielded the top ACC of 64.4% and the top AUC of 67.5% (Table 2).

Conventional SRC We then evaluated the performance of the conventional SRC method. We utilized multiple undersampling factors to address convergence to infeasible solutions mostly caused by linearly dependent vectors that yielded different classes. In Table 3, we show results from the top performing experiments producing 59.2% classification accuracy for resampling of 1/20, corresponding to feature dimensionality of 400 using leave-one-out cross-validation. We also applied conventional SRC to the texture feature set produced in section

Table 1 Classification performance for bone characterization using individual texture-based classifiers, or their ensembles, as denoted by the block size

Method	Feat. Sel.	Block Side	TPR (%)	TNR (%)	ACC (%)	AUC (%)
NB	CFS-GA	400	63.2	64.4	63.8	67.3
		100	38.4	64.8	51.7	54.6
		50	47.4	54.7	52.3	48.7
		25	46.9	65.9	51.7	55.2
BN	CFS-GA	400	66.7	62.1	64.4	70.4
		100	50.7	65.4	59.2	61.5
		50	37.9	55.2	46.6	50.2
		25	52.4	54.6	54.0	46.0
Bagging	CFS-GA	400	70.1	65.5	67.8	65.0
		100	44.6	57.3	50.6	52.1
		50	57.8	58.3	58.1	57.4
		25	46.1	55.1	51.2	53.4
RF	CFS-GA	400	67.8	65.5	66.7	68.2
		100	40.9	61.6	51.2	48.9
		50	45.0	51.1	48.3	50.0
		25	46.8	53.7	50.6	50.7
NB	CFS-BF	400	71.3	57.5	64.4	70.9
		100	43.9	59.8	52.3	52.0
		50	45.7	52.3	50.6	48.0
		25	48.5	53.7	51.7	49.6
BN	CFS-BF	400	64.4	66.7	65.5	69.9
		100	37.9	60.9	49.4	49.8
		50	40.2	46.0	43.1	44.9
		25	48.8	53.4	52.3	49.9
Bagging	CFS-BF	400	66.6	67.8	67.2	70.5
		100	50.0	63.3	56.9	52.6
		50	46.7	53.5	50.6	54.1
		25	58.6	51.0	54.0	50.4
RF	CFS-BF	400	60.9	67.8	64.4	68.4
		100	46.1	64.7	55.2	52.6
		50	48.0	55.6	52.3	52.2
		25	40.9	44.7	43.1	43.4

Table 2 Classification performance for bone characterization using Bag of Keypoints and deep learning techniques

Method	TPR (%)	TNR (%)	ACC (%)	AUC (%)
Bag of Keypoints	98.6	100	99.3	100
AlexNet	65.5	57.5	61.5	63.1
GoogleNet	64.4	54.0	59.2	65.6
Resnet18	80.5	48.3	64.4	67.5
Inceptionv3	69.0	51.7	60.3	66.5

“Nonparse Classification Techniques: Texture-Based, Patch-Based, and Deep Learning” and the classification accuracy was 71.7%.

Integrative Sparse Classification Here we report the performance of our block-based ensembles of

sparse classifiers. We employed block sizes ranging from 100×100 pixels to 10×10 pixels to observe the impact of this variable on the classification performance. We repeated these experiments using the BBMAP and BBLL decision functions in this

Table 3 Classification performance for bone characterization using conventional sparse classifiers and ensembles of block-based sparse classifiers. The block size in the first

Method	Block Side	TPR (%)	TNR (%)	ACC (%)	AUC (%)
BBMAP	400 (samp. 1/4)	55.2	54.0	54.6	58.4 4
	400 (samp. 1/20)	57.5	60.9	59.2	63.4
	100	65.5	67.8	66.7	71.4
	50	93.1	81.6	87.4	91.3
	25	100	100	100	100
	10	100	100	100	100
	Mean ± Std	89.7 ± 16.4	87.4 ± 15.7	88.5 ± 15.7	90.7 ± 13.5
BBLL ($\tau_{LLS}^* = 0$)	400 (samp. 1/4)	55.2	54.0	54.6	58.4
	400 (samp. 1/20)	57.5	60.9	59.2	63.4
	100	85.1	82.8	83.9	87.7
	50	98.6	90.8	94.8	97.3
	25	100	100	100	100
	10	100	100	100	100
	Mean ± Std	95.9 ± 7.2	93.4 ± 8.3	94.7 ± 7.6	96.3 ± 5.8

setting. We show our leave-one-out crossvalidation performance in Table 3. The experiment with block size 25×25 pixels that led to 256 classifiers performed the best classification of 100% by the BBMAP and BBLL techniques. These results imply 9.5% improvement of our method over the traditional SRC method. The block size with 10×10 also produced 100% accuracy and 100% AUC. In addition, we estimated the statistical significance of the differences between the AUC values of BBLL with optimized threshold τ_{LLS}^* and BBMAP by applying DeLong’s statistical test between the ROCs produced by BBMAP and BBLL. The p-values for block sizes of 100×100 , 50×50 , 25×25 and 10×10 were 0.47, 0.66, 0, and 0 respectively, suggesting significant differences for block sizes of 100×100 , 50×50 , 25×25 , and 10×10 . BBLL achieved the top AUC for 25×25 and 10×10 block size that was also found to be significantly different from the corresponding BBMAP result.

Discussion

The cross-validation experiments indicate that BBMAP and BBLL produce better separation than texture-based, BoK, deep neural networks, and SRC. Texture-based techniques may

two rows implies no block decomposition (as in conventional SRC)

achieve moderate classification rates. The feature computation and selection stages are key factors for improving separation accuracy. Smaller block sizes do not improve texture-based classification, because smaller sizes reduce the amount of information represented by the texture descriptors.

Bag of Keypoints, which shares some similarities with sparse representation methods, showed potential for exceptional results. This method computes patch-based feature vectors that are used for training and testing. SVM classification enables the estimation of nonlinear discriminant functions. On the other hand, deep learning methods did not perform very well, mainly because of the limited size of the training set. These methods have potential to outperform their conventional feature-based machine learning counterparts, if we are able to provide a big number of informative labeled samples to train the networks.

The results in Table 3 suggest that the block-based approach finds more accurate sparse solutions than the conventional SRC approach and improves the classifier performance. A reason for the improved group separation may be that the block-based ensemble technique employs multiple learners of over-complete dictionaries that are more amenable to sparse coding and representation. Between the block-based decision

functions, BBLL yielded higher classification rates for larger block sizes, because it accounts for estimation bias. Although both BBMAP and BBLL achieved perfect separation for small blocks, their performance drops when the number of training samples increases and the number of test samples decreases. We expect that effective dictionary learning will help to improve the generalization capability of these methods.

From the above studies and results, we conclude that machine learning-based methods using digital radiographs have the potential to assist the noninvasive diagnosis of osteoporosis, or the identification of high fracture risk. A meaningful element of such approaches would be to use the concept of signal sparsity to generate representations of the bone structure.

References

1. Bartl R, Frisch B. Osteoporosis: diagnosis, prevention, therapy. Springer Science & Business Media; Springer Berlin Heidelberg, 2009.
2. Hough S. Fast and slow bone losers. Relevance to the management of osteoporosis. *Drugs Aging*. 1998;12 (Suppl. 1):1–7. Available from: <https://doi.org/10.2165/00002512-199812001-00001>
3. Macintyre NJ, Lorbergs AL. Imaging-based methods for non-invasive assessment of bone properties influenced by mechanical loading. *Physiother Can*. 2012;64(2):202–15.
4. Martin-Badosa E, Elmoutaouakkil A, Nuzzo S, Amblard D, Vico L, Peyrin F. A method for the automatic characterization of bone architecture in 3D mice microtomographic images. *Comput Med Imaging Graph*. 2003;27(6):447–58.
5. Yger F. Challenge IEEE-ISBI/TCB: application of Covariance matrices and wavelet marginals. 2014; abs/1410.2663. Available from: <http://arxiv.org/abs/1410.2663>
6. Jennane R, Harba R, Lemineur G, Bretteil S, Estrade A, Benhamou CL. Estimation of the 3D self-similarity parameter of trabecular bone from its 2D projection. *Med Image Anal*. 2007;11(1):91–8. Available from: <http://www.sciencedirect.com/science/article/pii/S136184150600082X>
7. Jennane R, Ohley WJ, Majumdar S, Lemineur G. Fractal analysis of bone X-ray tomographic microscopy projections. *IEEE Trans Med Imaging*. 2001; 20(5):443–9.
8. Sonka M, Hlavac V, Boyle R. Image processing, analysis, and machine vision. Chapman & Hall Computing - London; New York, 2014.
9. Haralick RM. Statistical and structural approaches to texture. *Proc IEEE*. 1979;67(5):786–804.
10. Zheng K, Makrogiannis S. Bone texture characterization for osteoporosis diagnosis using digital radiography. In: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2016. p. 1034–1037.
11. Pentland AP. Fractal-based description of natural scenes. *IEEE Trans Pattern Anal Mach Intell*. 1984; PAMI-6(6):661–74.
12. Costa AF, Humpire-Mamani G, Traina AJM. An Efficient Algorithm for Fractal Analysis of Textures. In: 2012 25th SIBGRAPI Conference on Graphics, Patterns and Images; 2012. p. 39–46.
13. Blanco S, Figliola A, Quiroga RQ, Rosso O, Serrano E. Time-frequency analysis of electroencephalogram series. III. Wavelet packets and information cost function. *Phys Rev E*. 1998;57(1):932.
14. Shapiro L, Stockman G. Computer vision. Upper Saddle River: Prentice-Hall; 2001.
15. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature*. 1986;323(6088):533.
16. Duda RO, Hart PE, Stork DG. Pattern classification, 2nd Ed. Wiley-Interscience; New York 2001.
17. Mitchell TM. Machine learning. McGraw Hill series in computer science. McGraw-Hill; 1997. Available from: <http://www.worldcat.org/oclc/61321007>
18. Csurka G, Dance CR, Fan L, Willamowski J, Bray C. Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV; 2004. p. 1–22.
19. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems. 2012. p. 1097–1105.
20. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Computer Vision and Pattern Recognition (CVPR). 2015. Available from: <http://arxiv.org/abs/1409.4842>
21. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016.
22. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, et al. Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE Trans Med Imaging*. 2016;35(5): 1299–312.
23. He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. In: Leibe B, Matas J, Sebe N, Welling M, editors. Computer Vision – ECCV 2016. Cham: Springer International Publishing; 2016. p. 630–45.
24. Shin H, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging*. 2016;35(5):1285–98.

25. Yang J, Wright J, Huang TS, Ma Y. Image super-resolution as sparse representation of raw image patches. In: Computer Vision and Pattern Recognition, 2008 CVPR 2008 IEEE Conference on. 2008. p. 1–8.
26. Figueiredo MAT, Nowak R, Wright S. Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. *IEEE J Sel Top Signal Process.* 2007;1(4):586–97.
27. Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y. Robust face recognition via sparse representation. *IEEE Trans Pattern Anal Mach Intell.* 2009;31(2):210–27.
28. Qiao L, Chen S, Tan X. Sparsity preserving projections with applications to face recognition. *Pattern Recogn.* 2010;43(1):331–41.
29. Aharon M, Elad M, Bruckstein A. K-SVD: an algorithm for designing over-complete dictionaries for sparse representation. *IEEE Trans Signal Process.* 2006;54(11):4311–22.
30. Mairal J, Elad M, Sapiro G. Sparse representation for color image restoration. *Trans Img Proc.* 2008;17(1): 53–69.
31. Zepeda J, Kijak E, Guillemot C. SIFT-based local image description using sparse representations. In: Multimedia Signal Processing, 2009. MMSP'09. IEEE International Workshop on. IEEE; 2009. p. 1–6.
32. Zhao W, Xu R, Hirano Y, Tachibana R, Kido S. A sparse representation based method to classify pulmonary patterns of diffuse lung diseases. *Comput Math Methods Med.* 2015;2015:567932. Available from: <https://doi.org/10.1155/2015/567932>
33. Donoho DL, Elad M. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proc Natl Acad Sci.* 2003;100(5): 2197–202.
34. Davis G, Mallat S, Avellaneda M. Adaptive greedy approximations. *Constr Approx.* 1997;13(1):57–98.
35. Donoho DL. For most large underdetermined systems of linear equations the minimal ℓ_1 -norm solution is also the sparsest solution. *Commun Pure Appl Math.* 2004; 59(6):797–829.
36. Pati YC, Rezaifar R, Krishnaprasad PS. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In: Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on. IEEE; 1993. p. 40–44.
37. Aharon M, Elad M. Sparse and redundant modeling of image content using an image-signature-dictionary. *SIAM J Imaging Sci.* 2008;1(3):228–47.
38. Alizadeh F, Goldfarb D. Second-order cone programming. *Math Program.* 2001;95:3–51.
39. Zheng K, Harris CE, Jennane R, Makrogiannis S. Integrative blockwise sparse analysis for tissue characterization and classification. *Artif Intell Med.* 2020;107: 101885. Available from: <http://www.sciencedirect.com/science/article/pii/S0933365719303239>
40. Hassouni ME, Tafraouti A, Toumi H, Lespessailles E, Jennane R. Fractional Brownian motion and Rao geodesic distance for bone X-ray image characterization. *IEEE J Biomed Health Inform.* 2017;21(5):1347–59.



Artificial Intelligence in Laboratory Medicine

57

Davide Brinati, Luca Ronzio, Federico Cabitza, and Giuseppe Banfi

Contents

Introduction	804
A Gentle Introduction: What Is Machine Learning?	805
A Brief Overview of Machine Learning Implementation in Laboratory Medicine	805
Machine Learning Models in Laboratory Medicine	806
Conclusion	810
References	811

Abstract

The purpose of this chapter is to present the main literary sources concerning machine learning (so artificial intelligence) application to laboratory medicine field, with the appropriate language, understandable also for clinical chemists, medical staff, and so on. This goal is

tackled with a systematic debate of the main research works, developed in this field, in which the clinical task is well-defined and a proper statistical or machine learning algorithm is implemented. The main results for each research work will be described to underline and prove the workability and the feasibility of machine learning to laboratory data and the reliability of machine learning models to solve diagnostic or prognostic tasks and then support the medical staff in its decision process.

To provide a better overview also for non-expert audience (in terms of computer and data science), a gentle introduction will be provided, regarding what machine learning is and what machine learning experts strive to accomplish in their daily work routine. Moreover, the advantages of the transition from expert systems to machine learning systems and the general benefit of switching toward

D. Brinati (✉)
IRCCS Istituto Ortopedico Galeazzi, Milano, Italy
e-mail: davide.brinati@grupposandonato.it

L. Ronzio · F. Cabitza
Università Vita e Salute San Raffaele, Milano, Italy
e-mail: federico.cabitza@unimib.it;
federico.cabitza@unimib.it

G. Banfi
IRCCS Istituto Ortopedico Galeazzi, Milano, Italy
Università Vita e Salute San Raffaele, Milano, Italy
e-mail: banfi.giuseppe@unisr.it

machine learning systems (in laboratory medicine field) will be discussed.

Keywords

Machine learning · Artificial intelligence ·

Laboratory medicine · Literary review

Introduction

This chapter will present and report the latest applications of artificial intelligence (AI) in laboratory medicine. In particular, machine learning (ML) is the AI subfield that is going to be considered. The reasons for this interest in ML techniques lie on their capability to reach human performances on several tasks and in some cases even outperform them.

Laboratory medicine is traditionally used to apply technology and innovation, including automation, robotization, and computer science [1]. Quality control was introduced in clinical laboratory 60 years ago; data automatic transfer from instrument to a laboratory information system was routinely introduced in the 1980s and the telematic release immediately later. Industrialization of procedures, in other terms, is a characteristic of total laboratory automation and integration between chemistry, immunology, hematology, and serology with computer science, and it is a regular and expected pathway both inside and outside the hospital. Laboratory medicine was a vanguard in informatics technology applied to medicine; thus it seems obvious that laboratory should lead the AI projects in clinical medicine. However, this approach is moderately applied in the real world, although the interest of professionals is very high.

Some simple machine learning approaches were introduced in the early phase of informatization and robotization of the clinical laboratory. For example, the rules for internal quality control, the most famous are the Westgard ones, were introduced into the software of standalone instruments or in total lab automation systems. The Bull algorithm could be also quoted, based on the weighted mean of series of

20 consecutive patient results, particularly applied to control hematology instruments, where some parameters are very stable, having low biological variability: mean corpuscular hemoglobin concentration, for example, is low only in frank and severe anemia, and in this case the values are eliminated from the algorithm, and high in rare pathological conditions, namely, spherocytosis and cold agglutinins.

Some additional examples of clinical laboratory involvement in an early state of machine-based data evaluation are the retest when a specimen is measured to control a value out of alert range, and the reflex, i.e., additional tests performed on a specimen when a value is out of a reference range, as typically done for thyroid parameters. Image identification was introduced many years ago in urinalysis instruments for the sediment evaluation, and also in hematology systems, where the integration between data evaluation and slide preparation is now a common procedure.

In light of recent high-profile articles and editorials in high-impact journals (e.g., [2, 3]), it appears that with the decline of so-called expert systems, ML has gained a place in medicine and captured the interest of medical researchers and practitioners in predictive methods within this subfield of computer science.

The automatization of classification, segmentation, or anomaly detection tasks for both diagnostic and prognostic purposes can lead to an improvement of the medical service provided by the hospital facilities since doctors and medical staff can rely on additional guidance, which draws its conclusions from the evidence of data. Besides, the potential for applying machine learning to laboratory data deserves more attention by the readership of this journal, as well as by physician-scientists who will want to take advantage of this new computer-based support in pathology and laboratory medicine.

The ever-wider use of ML in clinical and basic medical research is reflected in the number of titles and abstracts of papers indexed on PubMed and published until 10 years ago (2006) as compared to the last 10 years (2007–2017), with a nearly tenfold increase from 1000 to slightly

more than 9000 articles in the past decade. In this chapter, the meaning of ML will be introduced in terms that physicians can easily grasp, and then the most recent applications of this computational approach to laboratory medicine will be surveyed.

The clinical laboratory is really ready for machine learning and artificial intelligence, also for integrating different specialized laboratories, having common methodologies, e.g., molecular biology, but different expertise, e.g., genetics, microbiology, and pathology.

A Gentle Introduction: What Is Machine Learning?

Put in very general terms, ML is about learning by machines. More specifically, ML is an umbrella term for diverse computational methods by which machines can incrementally build an accurate data model according to a measure of how well the model supports a given task, which in medicine is usually of discriminative nature, i.e., classification, clustering, or regressive. Classification, as the word says, regards the identification, for each record given to the model as input, of its target class, or correct category; on the other hand, regression regards the estimation of the correct value of a continuous variable. Lastly, clustering is an approach which allows to group different instances together, by associating them to groups (i.e., clusters) on the basis of their similarities. So, there are three players: *model*, *task*, and *measure of performances*. In an ML context, by model, we refer to the functional representation of a data set, i.e., the representation of any mapping that can be drawn to bind portions of the data set to a particular value on a specific measurement scale. The scale is either nominal or ordinal in a discriminative classification task, and the value is usually a label. In a discriminative regression task, on the other hand, the scale is either an interval or a ratio, and the predicted value is a number indicating some quantity, e.g., creatinine levels. Most of the ML models described in the medical literature so far regard functional mapping between a set of values, likely associated with a single clinical case, and a single category (e.g., yes/no or one

class out of a taxonomy) in order to support either a diagnosis or a prognosis.

The ML goal is to improve the model performance on a specific task by taking the measure of performance as a loss function to decrease (or increase, it depends on the task).

To illustrate, let's call this set of values x : in a prognostic decision task, the model is applied to answer questions like "does x represent (or values pertaining to) a patient who is affected by a certain disease or not?" In a supervised ML context, the data are usually data sets that describe different cases along various dimensions or attributes, called features, and that human experts have already associated with "correct" values, which we shall call " y ." Therefore, in the very concise terms that data scientists love, ML models work with functions like $y = f(x)$: the value of this function lies in its capability to yield the correct " y " also for some " x " that has not previously been classified by a human expert, thus providing aid in the classification task. Data scientists program machines, called "learners," to optimize a given model (i.e., make it more accurate in predicting y on the basis of a given unknown x taken from the target population) by autonomously and iteratively tweaking its parameters until no further improvement can or should be achieved: the process that the learners apply to a part of the available data, the "training data," without the direct intervention of human programmers, is called learning, in this case, machine learning.

A Brief Overview of Machine Learning Implementation in Laboratory Medicine

Machine learning is widely used in biochemical development and for evaluating and interpreting data in genomics, transcriptomics, and proteomics pathways (e.g., [4, 5]), whereas in clinical laboratory medicine, it has been applied to classical biomarker testing of biological materials. Several so-called expert systems have been recently described, patented, and commercialized for clinical laboratory purposes. Designed to evaluate specific data in hematology, urinalysis, or clinical

chemistry, they are traditionally based on a pre-defined decision tree encompassing logic rules and checks to exclude diagnostic hypotheses or define them or suggest further analysis to complete the diagnosis and support decision-making. By contrast, ML is a completely different approach, where “rules” are learned by the machine. More often than not, speaking of explicit rules is inappropriate, as the prediction is somehow hidden in the model’s nonlinear parameters that bend the decision boundaries around the data. Therefore, it’s clear that the application of ML in laboratory medicine should be supported as a means to enhance laboratory organization and expand the core skill set of laboratory experts, within a broader process of change and innovation.

There are several reasons supporting the previous sentence. First, laboratories are a major part of today’s healthcare systems. However, despite high throughput with low turnaround times, the capacity to screen data for results of special interest has decreased, and few tests are directly diagnostic [6]. Second, technological advances have enabled the integration of expert system capabilities and software applications, including auto-analyzers and modules of laboratory information systems (LIS) [7]. Since this kind of support is usually based on dichotomous thresholds or rigid mutual exclusion of data, it can be difficult if not impossible to obtain precise or personalized results [8], suggesting an obvious margin for improvement. Third, because patients can now immediately access their laboratory test results by downloading them from the Web portal of their diagnostic provider, there is an increasing demand for meaningful, possibly personalized reference limits and the need to interpret precision asterisks [6], the conventional signs indicating abnormal or borderline values. Finally, with the convergence of smartphones and innovative biosensors based on microfluidics and microelectronics, the vision of the lab-on-a-chip (LOC) and related models for laboratory medicine has opened chances [9].

In this context, apomediation refers to progressive disintermediation whereby traditional intermediaries, such as healthcare professionals who

give “relevant” information to their patients, are functionally replaced by apomediations, i.e., network/group/collaborative filtering processes [10]. ML systems can be seen as new and “cleverer” apomediations that act as gap fillers that analyze the increasing amount of diagnostic data a patient can access without intervention by a general practitioner or laboratory specialist and then assign the patient to a specialist only in case of likely positive or anomalous results. This can be done by factoring together the diverse phenotypic attributes of a patient (i.e., in addition to body mass index [BMI], age, gender, and ethnicity) or, better yet, of the patient’s history of past basal values associated with a healthy condition. In this case, the very notion of reference limits would change, and ML, by leveraging and improving other statistical approaches, could help limit the misinterpretation of values outside of reference limits or of apparently normal data but also diagnostic for some conditions.

Some imagine a ML-based clinical decision support that, by predicting correlated test results and improving the diagnostic value of multi-analyte sets of test results, could help to reduce unnecessary laboratory testing [11] and, hence, lower healthcare costs, which are budgeted to total \$5 billion yearly in the United States alone [12]. Finally, the growing number of available and affordable types of diagnostic tests has produced an unprecedented complexity of data interpretation and integration that calls for novel management technologies.

Machine Learning Models in Laboratory Medicine

In this section, several research works are reported, showing the potential of ML models to address the previously mentioned challenges in laboratory medicine.

Lin et al. [13] mined concepts from clinical narratives and lab values gathered from electronic medical records to automatically detect rheumatoid arthritis. After experimenting with a range of ML algorithms, they found that the linear kernel support vector machines performed best, with an

AUROC curve of 0.83, after which also inflammatory markers were considered (an increase of 6% as compared to no laboratory test).

Razavian et al. [14] collected administrative claims, pharmacy records, healthcare utilization, and lab test results of 4.1 million individuals between 2005 and 2009 to evaluate a prediction ML model for type 2 diabetes. Among the different variables associated with the development of diabetes, high aminotransferase (ALT) concentrations were associated with the highest odds ratio. Among the lab tests, the best prediction variables were glycated hemoglobin (HbA1c), glucose, high-density lipoprotein cholesterol, carbon dioxide, and glomerular filtration rate (GFR). The authors observed that the study also showed how administrative data can be a powerful tool for population health management and clinical hypothesis generation for risk factor discovery and that these data can help guide interventions in at-risk populations.

In a study involving 757 patients, Nelson et al. [15] applied logistic regression and an ML model they called “a relevance vector machine” and found that creatinine level was a clear predictor of outcome in traumatic brain injury, whereas glucose, albumin, and osmolality levels were predictors depending on the model used.

Diri and Albayrak [16] evaluated the performance of four classifiers applied to the data from five lab tests for thyroid dysfunction, distinguishing between a diagnosis of euthyroidism, hypothyroidism, and hyperthyroidism. The tests were as follows: T3-resin uptake test, total serum thyroxine, total serum triiodothyronine, basal thyroid-stimulating hormone (TSH), and maximal absolute difference of TSH value after injection of 200 μ g of the thyrotropin-releasing hormone as compared to the basal value. The authors also used cobwebs to visualize classifier performance when the data had more than two classes. A Bayesian classifier showed the best overall performance, with an average accuracy of 96%.

Given the complicated characteristics of warfarin, Liu et al. [17] used two well-known lab tests, alanine aminotransferase (ALT), and serum creatinine (SCr), in combination with data about warfarin dose, gender, age and weight, to build a

classification model that could predict adequate or inadequate warfarin therapy and minimize the odds of drug-to-drug interactions. In an analysis of 377 inpatients, they compared the performance of seven classification techniques and found that C4.5 decision tree and random forest scored best and predicted the adequacy of warfarin “more accurately than does the clinical physicians’ subjective decision.” This result, the authors claimed, showed the importance of making the best use of lab test results in clinical practice, especially in virtue of the relative simplicity and low cost of collecting accurate lab data.

Putin et al. [18] applied ML to laboratory parameters to predict chronological age via an ensemble of 21 deep neural networks developed and applied to more than 50,000 samples. They found that albumin concentration, followed by glucose, best identified chronological age. The ensemble identified five markers (albumin, glucose, alkaline phosphatase [ALP], urea, and erythrocytes) as the most valuable for predicting subject age.

Dermici et al. [19] used a commercial software program to train an ANN for application in the central laboratory of a large university hospital for the efficient, rapid, and reliable evaluation of biochemical test results. The ANN was applied to more than 250,000 samples to evaluate a set of routine parameters (sodium, potassium, calcium, magnesium, glucose, uric acid, chloride, urea, creatinine, aspartate aminotransferase, ALT, gamma-glutamyl transferase [GGT], ALP). Evaluation by ANN was compared with evaluation by seven pathologists of different expertise. The sensitivity of the model was 91% and specificity was 100%, with a K-score of 0.95. The K-score analysis revealed that five out of seven pathologists gave very high agreement scores in the evaluation of model judgment (0.81–1.00). When a reassessment of the specialists’ decision was requested, after comparison with the ANN evaluation, the pathologists changed their reports significantly in many cases, so as to increase agreement between the human and the automatically generated report. The time between receipt of the data and release of the reports was clearly lower in the case of ANN. The authors concluded

that a decrease in time and related costs, at similar quality and appropriateness levels, can be expected from the introduction of similarly accurate automatic supports.

Yuan et al. [20] built and evaluated three classifiers based on supervised ML methods to discriminate between positive and negative urine samples. Based on a classification and regression tree (CART), the model showed the best results on the test set, with a sensitivity of 86.0%, a specificity of 98.0%, an AUC of 94.3%, and an overall accuracy of 95.6%. The results implied that ML is a valuable method to construct classifiers for urine microscopic review rules that can supplement other reported microscopic review rules.

Advocates of ML methods affirm that ML may be useful for prognosis, i.e., predicting disease evolution and progression; early detection, when a disease is still in its early asymptomatic stage; and primary prevention to reduce the risk of development of the disease. The predictive approach has been based on regression models, e.g., the logistic model to predict 30-day mortality risk for patients with ST-segment elevation myocardial infarction (STEMI), the Weibull model for the SCORE (systematic coronary risk evaluation) model, and a Cox model applied to the Framingham Risk Score for cardiovascular diseases.

Goldstein et al. [21] described an ML method for cardiovascular risk prediction that was trained with the data of 1944 patients with a primary diagnosis of acute myocardial infarction. The authors used 13 lab parameters measured in at least 80% of the patients (calcium, carbon dioxide, creatinine, creatine kinase-MB, hemoglobin, glucose, mean corpuscular volume, mean corpuscular hemoglobin concentration, platelets, potassium, red cell distribution width [RDW], sodium, leukocytes) and calculated the median and the minimal and maximal values of these parameters to obtain 43 predictor variables of hospital mortality. The ML model trained on this data set showed that there is a nonlinear relationship between calcium and hemoglobin and post-infarction mortality. The authors employed five ML approaches to build models with different characteristics and performance: the variables were similarly relevant, and the models detected

the high impact of carbon dioxide (minimum value), calcium (all measures), hemoglobin (median value), potassium (all measures), and leukocytes (maximum value).

An epidemiological expert group reported on examples of ML applied to biochemical and hematological tests, including the demonstration of a relationship between GGT and liver function tests (ALP, albumin, lactate dehydrogenase, and aminotransferase), enhanced prediction of hepatitis B and C through the use of hepatitis C virus, and the correlation between RDW and hemoglobin in anemia diagnosis [22].

Somnay et al. [23] used serum levels of preoperative calcium, phosphate, parathyroid hormone, vitamin D, and creatinine as predictors of primary hyperparathyroidism in a sample of 11,830 patients. Among the ML algorithms tested, the Bayesian network models proved most accurate, correctly classifying 95% of all primary hyperparathyroidism patients (AUROC 0.99). Interestingly, excluding parathyroid hormone from the model did not substantially decrease its accuracy. The study concluded that ML can accurately diagnose primary hyperparathyroidism without human input even in cases of mild disease.

Luo et al. [24] investigated the utility of automated clinical decision support to predict test results using the results from other tests. As a proof of concept, they showed that ML models based on patient demographics (age and gender) and results of other lab tests (each collection had a median of 23 of the 40 tests) can discriminate normal from abnormal ferritin results with a high degree of accuracy (AUC 0.97, held-out test data) and even predict numerical results for ferritin (by regression) with moderate accuracy. They also reported that predicted ferritin results could better reflect underlying iron status than measured ferritin in some cases. Their results were shared by other studies, like that of Waljee et al. [25], who found that the miss Forest model outperforms other methods for imputing missing laboratory results.

Chen et al. [26] developed an ML model for predicting changes in GFR in Chinese patients with type 2 diabetes. Because current GFR equations (Cockcroft and Gault, Modification of Diet

in Renal Disease, Chronic Kidney Disease Epidemiology Collaboration) are known to be inaccurate in persons with diabetes, a model including sex, age, serum creatinine, and BMI provides an optimal modification of these equations in such patients.

Another interesting example of ML models that estimate the effectiveness and efficacy of well-established lab tests is the evaluation of the role of tumor markers in cancer diagnosis in asymptomatic subjects. Tumor marker testing is currently suggested for diagnostic assessment and especially during follow-up after chemotherapy but not during screening. Multiple tumor marker applications could be employed even in the screening phase, with the probabilistic power of a group of molecules reaching the proper edge, sufficient to identify asymptomatic disease. This approach is now supported by new metabolomics and proteomic approaches. Surinova et al. [27] found that five proteins not routinely measured in laboratories were selected by an ANN from among approximately 300 secreted and cell-surface candidate glycoproteins, which could represent a panel for the early diagnosis of colorectal cancer before clinical symptoms appear. Classical tumor markers, however, even when grouped by increasing sensitivity and specificity, are not useful for cancer screening in apparently healthy subjects, as reported by Wang et al. [28] who implemented an ML method to study its diagnostic power in screening with the tumor markers AFP, CEA, CA 19.9, CYFRA 21.1, and SCC, in addition to PSA for men and CA 15.3 and CA 125 for women. Evaluation of tumor marker screening in approximately 21,000 individuals showed an inadequate positive prediction value, a reduction in absolute risk and an increase in absolute risk. The authors concluded that combined tests should not be proposed for cancer screening.

Brinati et al. [29] propose ML models to improve the blood bag allocations and consequently save costs associated with patients' blood management system (PBM) of an important hospital facility. This study aims to demonstrate that ML models can improve the current hospital PBM strategy, based on a simple heuristic method

(threshold on hemoglobin value equal to 12.9 g/dl; if lower, the patient is treated as transfused), also in terms of cost-savings. The researchers use medical information and anographical personal data about 11,814 hospitalizations, attributable to 4593 patients, from January 2019 to May 2019. Each instance was provided with the following attributes: *age*, *gender*, *hemoglobin HGB (g/dl)*, *type of operation*, and *operation urgency*. The study presents ML models able to use such information to predict the transfusion risk and then treat the patients predicted as transfused with drugs, such as ferritin, in order to mitigate the transfusion risk. Several ML modes were trained and evaluated through nested cross-validation, which also allows hyperparameter tuning. Model evaluation has been focused on precision, recall, F1-score, and area under the ROC curve (AUC) as measures of performances, due to the class imbalance of the target variable. The best performing model is random forest that outperforms the current PBM strategy in terms of precision (or positive predictive value), F1-score, AUC, and cost per patient (computed with the associated cost for misclassification), with a net saving of about 30 euros per patient. This study represents an example of ML application aiming to improve not only the medical process but also the management of an important resource such as blood bags and its financial aspect.

Another interesting implementation of ML models for a diagnostic task is given by Brinati et al. [30], where ML techniques are employed for the detection of COVID-19 infection, using as input routine blood test values (hematochemical test). This study shows the workability and clinical soundness of using blood test analysis and ML as an alternative to rRT-PCR for identifying COVID-19-positive patients. This is especially beneficial in those countries, like developing ones, where rRT-PCR reagents and specialized laboratories are often missing. The authors of this work made available a web-based tool for clinical reference and evaluation, where everyone can upload his blood test data and look at the ML model response. The tool is available at this link: <https://covid19-blood-ml.herokuapp.com>. Data about 279 cases, randomly extracted from patients

admitted to the hospital from February 2020 to March 2020, was used for this study. Information contained in the dataset regards patient's age; gender; several routine blood test values such as alanine transaminase (ALT), aspartate transaminase (AST), white blood count (WBC), lactate dehydrogenase (LDH), gamma glutamyl transferase (GGT), C-reactive protein (CRP), etc.; and finally the result of the RT-PCR test for COVID-19. The ML models implemented and validated for this task are decision tree, extremely randomized tree, K-nearest neighbors, logistic regression, naive Bayes, random forest, and support vector machines. The model selection phase has been performed via nested cross-validation which also provides the best hyperparameters. The measures of performance adopted to evaluate models are accuracy, balanced accuracy (computed as the average of recall obtained on each class), precision (or positive predictive value), recall (or sensitivity), specificity (or true negative rate), and area under the ROC curve (AUC). Authors elected accuracy and recall as the main quality metrics since false negatives (i.e., patients positive to COVID-19 which are, however, predicted as negative and possibly let go home) are more deleterious than false positives in this screening task. The best model turned out to be random forest that records on the test set (20% of the initial instances) an accuracy equal to 82%, recall = 92%, precision = 83%, specificity = 65%, and AUC = 84%. The random forest also provides a feature importance rank, very useful for model interpretability. The feature importances were computed by estimating, for each feature, the total normalized reduction, across the decision trees in the trained random forest, to the variance of the target feature. The five most important features for the prediction of COVID-19 infection are AST, lymphocytes, LDH, CRP, and WBC. This chapter shows the feasibility of an alternative COVID-19 test (machine learning based) that works with hematochemical values as input.

Lastly, ML may be implemented to identify reference limits for lab parameters. Reference intervals should be determined for each specific test by the laboratory, considering the method used, the preanalytical phase, and the type of

population accessing the diagnostic service. However, because setting reference ranges is difficult, expensive, and time-consuming [31], reference ranges are generally collected from the literature or adopted from those suggested by the laboratory test manufacturers. The methods described in the specialist literature are usually based on traditional, descriptive statistical approaches and used to obtain reference limits directly from laboratory data. This is feasible, especially when a large amount of outpatient data is available or when the population is currently known as healthy or has a low prevalence of diseases.

Conclusion

The potential employment of machine learning models to laboratory data is relevant but not yet fully achieved. Although it is reasonable to expect that as machine learning techniques become better known they will be implemented to reduce costs, support clinical decision-making, and improve outcomes, further study is needed to understand whether and how the best machine learning practices can be advantageously transferred to laboratory medicine from other areas that pioneered this computational approach, like cardiology, oncology, and radiology, to tap into the related opportunities and strengths and avoid threats and weaknesses [32].

In the near future, ML will aid the pathologist and the clinical chemist to elaborate big amounts of data, and finally to make decisions, or, better, to suggest the best decision to clinicians. Definitely, ML will not substitute the experts, as predicted some years ago for radiologists. Actually, the current scientific scenario confirms that ML is superior when the number of facts (images) is very high and their interpretation is defined, but, when the incidence is low and the interpretation is difficult, there is no evidence toward the ML use only. Thus, in clinical laboratory, ML could aid in the diagnosis using the current parameters (high number of tests, defined thresholds, clear symptoms) and also could aid to elaborate, resume, and validate the big amount of data (e.g., genomics, miRNA, vitamin D, which can be validated by experts only for small

cohorts of patients) to define a possible link with symptoms or disease.

The spread of knowledge about ML (and AI), in the field of clinical chemistry, is a relevant topic and for its pursuit is essential to create a link between the clinical chemist and the expert on computer science and data science, to build a system which can be really effective for laboratory professionals and their meaningful use and safe adoption.

References

1. Badrick T, Banfi G, Bietenbeck A, Cervinski MA, Loh TP, Sikaris K. Machine learning for clinical chemists. *Clin Chem*. 2019;65(11):1350–6.
2. Darcy AM, Louie AK, Roberts LW. Machine learning and the profession of medicine. *J Am Med Assoc*. 2016;315:551–2.
3. Obermeyer Z, Emanuel EJ. Predicting the future – big data, machine learning, and clinical medicine. *N Engl J Med*. 2016;375:1216.
4. Camaggi CM, Zavatto E, Gramantieri L, Camaggi V, Strocchi E, Righini R, et al. Serum albumin-bound proteomic signature for early detection and staging of hepatocarcinoma: sample variability and data classification. *Clin Chem Lab Med*. 2010;48:1319–26.
5. Madabhushi A, Doyle S, Lee G, Basavanhally A, Monaco J, Masters S, et al. Integrated diagnostics: a conceptual framework with examples. *Clin Chem Lab Med*. 2010;48:989–98.
6. Horowitz GL. The power of asterisks. *Clin Chem*. 2015;61:1009–11.
7. Connelly DP. Embedding expert systems in laboratory information systems. *Am J Clin Pathol*. 1990;94 (4 Suppl 1):S7–14.
8. Lippi G, Bassi A, Bovo C. The future of laboratory medicine in the era of precision medicine. *J Lab Precis Med*. 2016;1:7.
9. Komatireddy R, Topol EJ. Medicine unplugged: the future of laboratory medicine. *Clin Chem*. 2012;58: 1644–7.
10. Eysenbach G. Medicine 2.0: social networking, collaboration, participation, apomediation, and openness. *J Med Internet Res*. 2008;10:e22.
11. Lindbury BA, Richardson AM, Badrick T. Assessment of machine learning techniques on large pathology sets to address assay redundancy in routine liver function test profiles. *Diagnosis*. 2015;2:41–51.
12. Jha AK, Chan DC, Ridgway AB, Franz C, Bates DW. Improving safety and eliminating redundant tests: cutting costs in U.S. hospitals. *Health Aff*. 2009;28:1475–84.
13. Lin C, Karlson EW, Canhao H, Miller TA, Dligach D, Chen PJ, et al. Automatic prediction of rheumatoid arthritis disease activity from the electronic medical records. *PLoS One*. 2013;8:e69932.
14. Razavian N, Blecker S, Schmidt AM, Smith-McLallen A, Nigam S, Sontag D. Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data*. 2015;3:277–8.
15. Nelson DW, Rudell A, MacCallum RM, Holst A, Wanecik M, Weitzberg E, et al. Multivariate outcome prediction in traumatic brain injury with focus on laboratory values. *J Neurotrauma*. 2012;29:2613–24.
16. Diri B, Albayrak S. Visualization and analysis of classifiers performance in multi-class medical data. *Expert Syst Appl*. 2008;34:628–34.
17. Liu KE, Lo CL, Hu YH. Improvement of adequate use of warfarin for the elderly using decision tree-based approaches. *Methods Inf Med*. 2014;53:47–53.
18. Putin E, Mamoshina P, Aliper A, Korzinkin M, Moskalev A, Kolosov A, et al. Deep biomarkers of human aging: application of deep neural networks to biomarker development. *Aging*. 2016;8:1021.
19. Demirci F, Akan P, Kume T, Sisman AR, Erbayraktar Z, Sevinc S. Artificial neural network approach in laboratory test reporting. *Am J Clin Pathol*. 2016;146:227–37.
20. Yuan C, Ming C, Chengjin H. UrineCART, a machine learning method for establishment of review rules based on UF-1000i flow cytometry and dipstick or reflectance photometer. *Clin Chem Lab Med*. 2012;50:2155–61.
21. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *Eur Heart J*. 2017;38:1805–14.
22. Richardson A, Signor BM, Lindbury BA, Badrick T. Clinical chemistry in higher dimensions: machine-learning and enhanced prediction from routine clinical chemistry data. *Clin Biochem*. 2016;49:1213–20.
23. Somnay YR, Craven M, McCoy KL, Carty SE, Wang TS, Greenberg CC, et al. Improving diagnostic recognition of primary hyperparathyroidism with machine learning. *Surgery*. 2017;161:1113–21.
24. Luo Y, Szolovits P, Dighe AS, Baron JM. Using machine learning to predict laboratory test results. *Am J Clin Pathol*. 2016;145:778–88.
25. Waljee AK, Mukherjee A, Singal AG, Zhang Y, Warren J, Balis U, et al. Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open*. 2013;3:e002847.
26. Chen J, Tang H, Lv L, Wang Y, Liu X, Lou T. Development and validation of new glomerular filtration rate predicting models for Chinese patients with type 2 diabetes. *J Transl Med*. 2015;13:300–17.
27. Surinova S, Choi M, Tao S, Schuffler PJ, Chang CY, Clough T, et al. Prediction of colorectal cancer diagnosis based on circulating plasma proteins. *EMBO Mol Med*. 2015;7:1166–78.
28. Wang HY, Hsieh CH, Wen CN, Wen YH, Chen CH, Lu JJ. Cancers screening in an asymptomatic population by using multiple tumour markers. *PLoS One*. 2016;11:e0158285.

29. Brinati D, Seveso A, Perazzo P, Banfi G, Cabitza F. Evaluation of cost-saving machine learning methods for patient blood management. In: 12th international conference on e-Health, 21–23 July 2020.
30. Brinati D, Campagner A, Ferrari D, Locatelli M, Banfi G, Cabitza F. Detection of COVID-19 infection from routine blood exams with machine learning: a feasibility study. *J Med Syst.* 2020;44(8):1–12.
31. Henny J, Vassault A, Boursier G, Vukasovic I, Mesko Brguljan P, Lohmander M, et al. Recommendation for the review of biological reference intervals in medical laboratories. *Clin Chem Lab Med.* 2016;54:1893–900.
32. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *J Am Med Assoc.* 2017;318:517–8.



Artificial Intelligence in Medicine (AIM) in Cardiovascular Disorders

58

Hisaki Makimoto

Contents

Introduction	814
Overview of AI Research on Cardiovascular Disorders	814
Research in General Cardiovascular Disorders	815
Research Targeting Ischemic Heart Diseases	815
AI in Noninvasive Evaluations	816
AI During Invasive Procedures	816
AI in Risk Assessments	816
Research Targeting Heart Failure	817
AI in Diagnosing Heart Failure	817
AI to Predict Prognosis of Heart Failure	817
Research Targeting Arrhythmias	818
AI to Diagnose Arrhythmias	818
AI to Monitor for Arrhythmias	819
AI to Identify Arrhythmias from Sources Other than ECGs	819
Discussion	819
References	820

Abstract

Cardiovascular disorders are one of the major causes of death in developed countries, and they form an important entity in clinical medicine and medical economics due to extended life expectancy. The diagnosis and treatment of cardiovascular diseases require multiple

complex laboratory tests and invasive procedures. Attempts to introduce artificial intelligence (AI) technology are underway in each stage of these processes. Electrocardiogram diagnosis is the most traditional and basic cardiovascular examination, and many diagnostic programs based on algorithms and machine learning have been developed so far. Introduction of AI has also been attempted in other noninvasive diagnostic tests, such as echocardiography and cardiac computed tomography and invasive examinations and treatments, such as coronary angiography and catheter interventions. With the growing popularity of

H. Makimoto (✉)
Arrhythmia Service, Division of Cardiology, Pulmonology and Vascular Medicine, Faculty of Medicine, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany

deep learning, AI programs are making progress in their detection and diagnostic abilities. Several research groups have aimed toward more accurate detection or more efficient treatment using AI. This chapter discusses previously reported AI research in the areas of general cardiovascular care, ischemic heart disease, heart failure, and arrhythmia.

Keywords

Cardiovascular diseases · Coronary artery disease · Heart failure · Arrhythmias

Introduction

Cardiovascular disorders are well-known major causes of death in developed countries. This trend will continue for a while due to extending life expectancy. Cardiovascular disorders remain significant in general medicine and health economics in such countries.

Diagnosing cardiovascular disorders require multilaterally specialized and complex processes to determine the fundamental cardiovascular pathologies underlying the various comorbidities or individual differences between patients.

After diagnosing them, the treatment of cardiovascular disorders also consists of complex steps. A number of cases require invasive procedures, such as catheter-based interventions, in which multiple examinations run continuously in parallel. Additionally, regarding the drugs prescribed in the hospital or after discharge, it is not easy to find the optimal dosage, particularly, in patients of advanced age because most of them have multiple underlying diseases and require a combination of drugs.

In current clinical situations, complicated diagnosis and treatment processes are accomplished according to medical standards, such as guidelines, and are based on the experience of each practitioner.

Artificial intelligence (AI) technology has been introduced into the field of cardiovascular medicine because it has the potential to recognize the information that human physicians cannot distinguish or even recognize, which could be an

advantage in the process of complicated diagnosis and treatment. It also suggests the possibility of an order-made diagnosis and treatment based on individual data. In this chapter, reports on AI technology are discussed with a focus on the diagnosis and treatment of cardiovascular disorders at hospitals.

Overview of AI Research on Cardiovascular Disorders

Most AI programs for cardiovascular disorders are based on machine learning. Recently, more groups have started to adopt deep learning using neural networks instead of predefined machine learning algorithm constructions.

The input data for building such AIs have included data of electrocardiogram (ECG), chest X-rays, echocardiography, cardiac computed tomography (CT), cardiovascular magnetic resonance imaging (CMR), and laboratory blood results.

The most common data source has been ECG, which records the electrical signals of the myocardium through electrodes on the body. In clinical practice, the signals are usually recorded using 12 leads and are described as two-dimensional (2D) waveforms for interpretation by physicians. In studies that used ECGs as input data for building AIs, various forms of ECG data were adopted, such as electrical raw data, 2D image data, or data recorded with reduced leads, such as single-lead or monitor ECG.

Chest X-rays, echocardiography, cardiac CT, and CMR are few other diagnostic imaging modalities. The image data, parameters, and diagnostic information from these investigations have also been utilized in building AI.

The amount of input data used for machine learning has increased from hundreds to millions of cases in some projects.

Studies targeting general cardiovascular disorders have been aimed at efficient diagnosis using each of these investigations. Their results suggest further possibilities in the subsequent diagnostic and treatment steps. There have also been projects that have targeted certain cardiac disorders, such as ischemic heart disease, heart failure, or

arrhythmia, which constitute a majority of cardiovascular disorders with an aim of early discovery and efficient treatment.

Research in General Cardiovascular Disorders

Studies on general cardiovascular disorders have mainly focused on major investigations in the current diagnostic processes.

ECG is a major traditional examination in cardiovascular disorders. Before the introduction of AI technology, computer programs had been developed to detect ECG abnormalities and disorders based on fixed algorithms [1]. These are commonly installed on ECG equipment currently used in clinical practice. Recently, several research groups have developed programs built using machine learning, including deep learning, to interpret ECG data. For automated ECG analysis, identification of the heartbeats in the ECG data is a fundamental step. A research group has previously developed a program based on K-Nearest-Neighbor (KNN) algorithm to recognize the heartbeats using electrical raw ECG signals [2].

Echocardiography is also a major diagnostic modality in the assessment of cardiac function and anatomy in cardiovascular disorders. It involves the emission of high-frequency ultrasonic waves toward the heart and visualization of its internal structures based on the reflected waves. Three-dimensional (3D) reconstruction and visualization of the heart is achieved by emitting the ultrasound waves from various angles. Therefore, echocardiography enables physicians in evaluating whether the myocardium is moving correctly, cardiac contraction per heartbeat, and structural defects, such as valve stenosis or regurgitation.

For the analysis of these video image data, it is essential to identify the angle of each image as well. The identification of this angle leads to an understanding of the observed cardiac anatomy. A program has been developed using deep learning to recognize the angle from which the echocardiographic image was obtained [3, 4]. Although the angles that can be distinguished were limited to a

few, these reports suggest the possibility of advanced features in analyzing the echocardiography data using AI.

AI has also been introduced in pharmaceutics. Warfarin is a commonly used traditional anticoagulant in cardiovascular disorders to prevent thrombus formation. It is widely used in patients following cardiac valve replacement or those with arrhythmias, such as atrial fibrillation. However, it is not easy to identify the optimum individual dose because of its several interactions with other medications and foods. A research group has attempted to predict the optimum dose of warfarin based on the clinical data of each patient using machine learning [5]. They have suggested the possibility of AI in efficient order-made medicine although their results did not seem sufficient for clinical and practical use.

There have also been studies on AI technology based on general data. One research group has developed a program to predict the risk of cardiovascular disorders using results of blood investigations [6]. Another group has reported a program based on transdermal optical imaging for the estimation of blood pressure using the image of faces recorded using smartphones [7]. These results could contribute to the prevention or early discovery of cardiovascular disorders.

Research Targeting Ischemic Heart Diseases

Ischemic heart disease is a clinical condition in which the blood flow in the coronary arteries is blocked at one or more sites. Consequently, oxygen cannot be supplied to the myocardium distal to the occluded sites, thus, leading to myocardial dysfunction. Myocardial infarction (MI) and angina pectoris are typical ischemic heart diseases with typical symptoms of chest pain and dyspnea. In simple terms, a coronary blood vessel is partially obstructed in angina pectoris and completely occluded in MI in which necrosis of the myocardium has begun. It has been reported that 15% of 30-year-old individuals will develop ischemic heart disease (MI or angina pectoris) in their lifetimes [8].

Most ischemic heart diseases are induced by atherosclerotic growth in the coronary arteries. The blood vessels, including coronary arteries, harden with age or due to other reasons, where the deposition of atherosclerotic plaques, such as cholesterol, progresses on the endovascular wall. When an atherosclerotic plaque suddenly ruptures and forms blood clots in the vessels, they can occlude the coronary arteries, thus, resulting in necrosis of the distal myocardium due to insufficient oxygen supply. Arteriosclerosis can occur due to high blood pressure, diabetes, renal dysfunction, or smoking and increase the risk of ischemic heart diseases. MI or angina pectoris occurs frequently due to physical burdens, such as strenuous exercise, psychological stress, or sudden fall in temperature.

If ischemic heart disease is suspected based on the onset and progression of clinical symptoms, the first diagnostic step is ECG. If ischemic heart disease is also suggested by ECG, echocardiography and cardiac CT are performed to determine the pathologic changes and to identify the sites of stenosis/occlusion as well as assess their severity. Furthermore, the degree of myocardial necrosis should be evaluated using blood sampling. If severe stenosis is noted, the site of stenosis should be dilated invasively using catheter interventions to restore the blood flow. The current guidelines recommend immediate catheterization in case of MI to avoid necrosis of the myocardium [9].

AI in Noninvasive Evaluations

There are some studies regarding AI technology aimed at the improved efficiency of noninvasive diagnostic and treatment processes.

Some groups have developed programs using machine learning to detect MI based on ECG [10–12]. One of those programs could judge MI and the site of stenosis in every heartbeat [10]. Other groups have reported programs built using deep learning based on electrical raw ECG data [11] or using ECG image data to recognize MI at least as accurately as human physicians [12].

There have also been efforts to introduce AI technology into cardiac CT assessments. It has

been reported that an algorithm built using machine learning can estimate the size of the ischemic area and the severity of stenosis using cardiac CT images [13]. By combining deep learning and other machine learning methods, the ischemic area can be estimated on cardiac CT images to identify those areas that require catheter interventions [14]. The severity of coronary artery stenosis is often assessed using an invasive index called fractional flow reserve (FFR), which functions as a criterion for further invasive catheter interventions. A study group has attempted to estimate FFR noninvasively based on cardiac CT data [15].

AI During Invasive Procedures

In coronary angiography (CAG), a contrast medium is injected during cardiac catheterization into the target coronary artery to assess the stenosis. A study group has developed a program using deep learning to recognize the branches of coronary arteries automatically and the shape of stenosis using CAG images [16]. To evaluate the state of the blood flow, it is important to assess CAG images that are synchronized with each heartbeat from various angles. An AI program has been reported to identify the image frames of the precise timing/phase in the cardiac contraction cycle from various angles [17].

For a detailed analysis for coronary stenosis, the instantaneous wave-free ration (iFR) is one of the modalities to assess hemodynamic appropriateness for catheter intervention. A group has reported that the severity of coronary artery stenosis and requirement for invasive catheter intervention could be well interpreted using machine learning [18].

Although several stages are yet to be achieved before such technologies can be used in clinical practice, they may be developed into an automatic analysis system of CAG images in the future.

AI in Risk Assessments

Patients who have experienced ischemic heart disease are predisposed to higher risks of subsequent cardiac arrest and mortality [19]. Therefore,

the prediction of their risks accurately is an important theme in order to mitigate these risks after the diagnosis of ischemic heart diseases.

An AI program can estimate mortality using machine learning based on cardiac CT and other clinical data [20]. Another group has also reported AI based on the random-forest regression model to predict the risk of rehospitalization due to heart failure, and cardiovascular mortality after catheter interventions [21]. The use of a multilayer perceptron has enabled the prediction of inhospital mortality and mortality over 1 year after discharge from the hospital under the diagnosis of acute MI more accurately than the risk stratification using the current clinical scoring system [22].

Research Targeting Heart Failure

Heart failure is a comprehensive clinical condition caused by dysfunction of the heart pump. Breathlessness, ankle edema, and fatigue are typical symptoms of heart failure. Left ventricular ejection fraction (LVEF) is a major index of cardiac function. LVEF in patients with heart failure can be in the normal range ($\geq 50\%$), midrange (40–49%), or highly reduced range (<40%), which are defined as heart failure with preserved ejection fraction (HFpEF), heart failure with midrange ejection fraction (HFmrEF), and heart failure with reduced ejection fraction (EFrEF), respectively [23]. The ejection fraction is assessed using imaging tests, such as echocardiography and CMR. The common causes of heart failure include MI, valvular disease, cardiomyopathy, and myocarditis.

People of advanced age are prone to develop heart failure. The morbidity in patients with heart failure is reported to increase, which corresponds to the extending average life expectancy [24, 25]. The Framingham Heart Study has reported that heart failure was observed at 0.8% in the 50s and 6.6% in the 80s [26]. The prognosis of patients with heart failure is poor. The mortality of patients hospitalized due to heart failure is three times higher than that in patients without heart failure [27].

AI in Diagnosing Heart Failure

HFrEF can be diagnosed based on significantly reduced cardiac function as assessed using echocardiography and cardiac MRI. If HFrEF is diagnosed, its causative cardiovascular diseases should be identified, which requires complicated professional evaluations. Several attempts have been made to introduce AI technology into this diagnostic process.

A program built using deep learning has been reported to interpret the echocardiography findings to not only assess the cardiac function but also suggest the causative diseases out of three possible choices [28].

In HFpEF, the apparent cardiac function appears to be normal, but symptoms of heart failure are noted. It is believed to be associated with insufficient relaxation of the left ventricular myocardium. The prevalence of HFpEF has increased in the last 20 years [29]. A study group has reported an AI program built using clustering to classify patients with HFpEF into three phenotypes [30], which could contribute in efficient risk stratification of their prognosis.

Both constrictive pericarditis and restrictive cardiomyopathy can result in HFpEF; however, it is not easy to distinguish them using echocardiography, especially, for inexperienced physicians. A research group has developed a program using cognitive machine learning to distinguish these two conditions based on the parameters of speckle tracking echocardiography [31].

AI to Predict Prognosis of Heart Failure

As mentioned above, the prognosis of patients with heart failure is poor, and it is important to predict their prognosis as precisely as possible in order to prevent life-threatening aggravations. The Seattle Heart Failure Model (SHFM) is a traditional scoring model used to predict such prognosis using multiple clinical data [32]. A study group has adopted logistic regression and random-forest methods to demonstrate the possibility of a better risk prediction with SHFM parameters using additional clinical data

[33]. Another group has reported a recurrent neural network (RNN) model to predict the risk of rehospitalization due to heart failure during 12 or 18 months of follow-up from the first onset of heart failure using the data of the diagnosis, drugs, and procedures [34].

Some research groups have applied machine learning in the general population not limited to patients with heart failure. A CNN model has been developed to detect severe cardiac dysfunction on ECG [35]. Interestingly, the individuals in whom cardiac dysfunction was screened as “positive” by the CNN but actual cardiac function was judged “not severely reduced” according to echocardiography demonstrated a four times higher risk of developing cardiac dysfunction in the future. This suggests that there could be signs on ECGs that human physicians cannot recognize at the moment.

Furthermore, a research group reported an AI program to predict the inhospital mortality in patients who underwent transcatheter aortic valve implantation (TAVI) [36]. This program judged mortality based on the given clinical information and demonstrated better predictions than the current scoring systems.

Research Targeting Arrhythmias

Arrhythmia is the general term for abnormal heartbeats, such as tachycardia, bradycardia, or extrasystole. Tachycardia is defined as rapid heart rate ($\geq 100/\text{min}$) and needs treatments in cases of symptoms, such as palpitations, shortness of breath, chest pain, dizziness, or syncope. Bradycardia is defined as slow heart rate (≤ 50 or $60/\text{min}$) and necessitates treatments in case of symptoms, such as shortness of breath or dizziness. Extrasystole is a premature contraction that occurs in the atrial or ventricular myocardium independent of the normal rhythm. The cause of these arrhythmias varies widely between patients; however, the prevalence of arrhythmias generally increases with age and increasing comorbidities.

During the normal heart rhythm (sinus rhythm), electrical signals are generated regularly at the sinus node in the right atrium, conducted to

the left atrium, and subsequently to the right and left ventricles through the atrioventricular node, which leads to myocardial contraction. If this process is disturbed, it can result in arrhythmias. For example, ischemic or valvular heart diseases can trigger arrhythmias where the electrical signals are often blocked because of the injured myocardium. Abnormal activities, such as extra signaling, may be enhanced in this injured area. Hereditary genetic disorders, electrolytic abnormalities, or side effects of drugs can also evoke arrhythmias.

Atrial fibrillation, ventricular tachycardia, and ventricular fibrillation are common arrhythmias that influence the clinical outcomes. During atrial fibrillation, the atrial myocardium trembles rapidly and finely ($>400/\text{min}$), which can result in a thrombus in the left atrium, thus, predisposing the patients to thromboembolic stroke. Ventricular tachycardia and fibrillation impair the heart’s ability to pump blood because of the rapid excitement of the myocardium in the ventricle (ventricular tachycardia) or because of the uncoordinated trembling of the ventricle (ventricular fibrillation), which increases the risk of sudden cardiac death.

Arrhythmias are diagnosed using ECG and should be treated according to the life-threatening risk and symptoms. Tachyarrhythmias can lead to lethal incidents, such as a stroke or sudden cardiac death; therefore, catheter intervention should/ could be considered to identify and eliminate the origin of arrhythmias using radiofrequency or cryoablation (catheter ablation). In the case of bradyarrhythmia that can possibly lead to lethal incidents or result in serious problems, such as syncope, pacemakers can be implanted to maintain sufficient heart rate.

AI to Diagnose Arrhythmias

ECGs are imperative in diagnosing arrhythmias. In this context, some research groups have reported on the detection of arrhythmias based on ECGs using deep learning.

Recently, a model was developed by a Chinese study group to derive rhythmic diagnoses using

12-lead-ECGs [37]. This model could choose one or more diagnoses from 21 choices of rhythm or conduction abnormalities. They gathered over 180.000 ECGs from more than 70.000 patients and utilized over 130.000 ECGs to train their CNN model using residual blocks. This model achieved a higher accuracy than that of 53 cardiologists, including experts.

Another group developed a CNN model using residual blocks to derive a diagnosis based on 12-lead-ECG from six choices of heart rhythm and conduction disorders [38]. They gathered over 2.300.000 ECGs from approximately 170.000 patients for this project.

ECGs must be recorded during the arrhythmia for a precise diagnosis; however, arrhythmias do not always appear while ECGs are being recorded. A model to solve this problem has been reported to identify potential patients with atrial fibrillation based on their sinus rhythm on ECG [39]. They used approximately 650.000 electrical raw 12-lead-ECG data from over 180.000 patients. Additionally, approximately 450.000 and 60.000 ECGs were used for training and validation, respectively. The remaining 130.000 ECGs were used for testing. They used a CNN architecture with residual blocks. This model would be useful not only for the early discovery of arrhythmias but also for the treatment strategy, for example, in patients with prior embolic stroke of undetermined source (ESUS). Patients with ESUS are divided according to the presence of atrial fibrillation because they require different therapeutic strategies that can otherwise lead to lethal incidents, such as restroke or intracranial bleeding. Identifying potential atrial fibrillation based on sinus rhythm on ECGs will help in developing appropriate medication strategies.

AI to Monitor for Arrhythmias

Some studies have focused on detecting arrhythmias and not the underlying diagnosis.

ECGs recorded through fewer electrodes are less useful for precise diagnoses but can be used to detect the heart rhythm. Fewer electrodes result in smaller ECG equipment and easier application in

mobile equipment, such as smart watches. Monitoring and early detection of arrhythmias in daily life can be highlighted with these small wearable devices. For such ECG monitoring using fewer leads, AI technology has been introduced toward efficient detection of arrhythmias. This technology will provide potential patients a chance for early treatment or secure the patients with cardiovascular disorders from their high-risk prognosis.

The functionality of a program based on algorithms to detect arrhythmias has been previously reported [40]. Using a machine learning methodology called echo state network, a group demonstrated the possibility of detecting ventricular arrhythmias efficiently [41]. A 33-layer CNN model to detect one of the 12 rhythm classes has also been reported [42]. Over 90.000 single-lead ECGs from more than 50.000 patients were utilized to train this model.

Various research groups have demonstrated that false arrhythmia alarms based on ECG monitoring in the intensive care units could be reduced using AI programs built using machine learning based on ECG data monitored with fewer electrodes combined with additional clinical data (arterial blood pressure and photoplethysmogram signals) [43, 44].

AI to Identify Arrhythmias from Sources Other than ECGs

There is an interesting report on arrhythmia detection using deep learning based on general data and not medical data [39]. In their presentation, the group reported that atrial fibrillation could be detected based on video images of the patients' face. This may provide potentially asymptomatic patients with the opportunity for early treatment. Further developments are expected.

Discussion

AI technology has demonstrated great potential toward more efficient diagnostic and treatment processes in cardiovascular disorders. Its recognition capability, which is superior to that of

humans, can be a strong advantage considering the complex evaluations in clinical practice. Most of the already accumulated clinical data are objective and easy to digitalize for building AI. According to reports, a large amount of data is not always necessary to build an AI, which suggests the possibility of model development suitable for minority groups or each individual.

However, AI technology is not yet mature enough for clinical use in the field of cardiovascular medicine. Each model reported so far could solve only simple questions, unlike human physicians who can process various clinical information, including subjective and nondigitalized data simultaneously. Some of the studies were at the prestage in the diagnostic process, for example, to judge the angle of images. Furthermore, we cannot judge whether the AI “learns” as appropriately as human medical students or residents. AI can be easily deceived by fake or unimportant information [45]. Further technical developments in AI are essential to guarantee quality and safety in clinical use.

Despite the aforementioned problems, we can still expect AI technology to overcome certain challenges. The current simple programs can be combined into providing more complicated and comprehensive judgments. The weakness in guaranteeing the learning process can be partially overcome using other ideas, such as visualization of the parameters of neural networks with a heatmap. The heatmap could help in visualizing the points on ECG that the AI focused on for its judgment by using the ECGs as image data and not electric raw data [12], which can help the AI “explain” its thought process.

In summary, political and legal agreements aside, AI can be useful in clinical situations, at least as an assistant, particularly, in the absence of professional equipment or expert physicians. The quality and safety can be guaranteed by using final decisions of human physicians. AI technology is still developing intensively and its technical limitations will be resolved in the future, which will help improve the trust in AI and integrate it into clinical settings. It is expected that AI technology will bring significant progress in the diagnostic and treatment processes in cardiovascular disorders.

References

- Willems JL, Abreu-Lima C, Arnaud P, van Bemmel JH, Brohet C, Degani R, Denis B, Gehring J, Graham I, van Herpen G, et al. The diagnostic performance of computer programs for the interpretation of electrocardiograms. *N Engl J Med.* 1991;325(25):1767–73. <https://doi.org/10.1056/NEJM19911219325203>.
- Saini I, Singh D, Khosla A. QRS detection using K-Nearest Neighbor algorithm (KNN) and evaluation on standard ECG databases. *J Adv Res.* 2013;4(4):331–44. <https://doi.org/10.1016/j.jare.2012.05.007>.
- Madani A, Arnaout R, Mofrad M, Arnaout R. Fast and accurate view classification of echocardiograms using deep learning. *NPJ Digit Med.* 2018;1:6. <https://doi.org/10.1038/s41746-017-0013-1>.
- Kusunose K, Haga A, Inoue M, Fukuda D, Yamada H, Sata M. Clinically feasible and accurate view classification of echocardiographic images using deep learning. *Biomol Ther.* 2020;10(5):665. <https://doi.org/10.3390/biom10050665>.
- Tao Y, Zhang Y, Jiang B. DBCSMOTE: a clustering-based oversampling technique for data-imbalanced warfarin dose prediction. *BMC Med Genet.* 2020;13 (Suppl 10):152. <https://doi.org/10.1186/s12920-020-00781-2>.
- Miranda E, Irwansyah E, Amelga AY, Maribondang MM, Salim M. Detection of cardiovascular disease risk's level for adults using naive bayes classifier. *Healthc Inform Res.* 2016;22(3):196–205. <https://doi.org/10.4258/hir.2016.22.3.196>.
- Luo H, Yang D, Barszczky A, Vempala N, Wei J, Wu SJ, Zheng PP, Fu G, Lee K, Feng ZP. Smartphone-based blood pressure measurement using transdermal optical imaging technology. *Circ Cardiovasc Imaging.* 2019;12(8):e008857. <https://doi.org/10.1161/CIRCIMAGING.119.008857>.
- Rapsomaniki E, Timmis A, George J, Pujades-Rodriguez M, Shah AD, Denaxas S, White IR, Caulfield MJ, Deanfield JE, Smeeth L, Williams B, Hingorani A, Hemingway H. Blood pressure and incidence of twelve cardiovascular diseases: lifetime risks, healthy life-years lost, and age-specific associations in 1.25 million people. *Lancet.* 2014;383(9932):1899–911. [https://doi.org/10.1016/S0140-6736\(14\)60685-1](https://doi.org/10.1016/S0140-6736(14)60685-1).
- Collet JP, Thiele H, Barbato E, Barthélémy O, Bauersachs J, Bhatt DL, Dendale P, Dorobantu M, Edvardsen T, Folliguet T, Gale CP, Gilard M, Jobs A, Jüni P, Lambrinou E, Lewis BS, Mehilli J, Meliga E, Merkely B, Mueller C, Roffi M, Rutten FH, Sibbing D, Siontis GCM, ESC Scientific Document Group. 2020 ESC Guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation. *Eur Heart J.* 2020. <https://doi.org/10.1093/eurheartj/ehaa575>.
- Arif M, Malagore IA, Afsar FA. Detection and localization of myocardial infarction using K-Nearest Neighbor classifier. *J Med Syst.* 2012;36(1):279–89. <https://doi.org/10.1007/s10916-010-9474-3>. Epub 25 Mar 2010

11. Strodtboff N, Strodtboff C. Detecting and interpreting myocardial infarction using fully convolutional neural networks. *Physiol Meas.* 2019;40(1):015001. <https://doi.org/10.1088/1361-6579/aaf34d>.
12. Makimoto H, Höckmann M, Lin T, Glöckner D, Gerguri S, Clasen L, Schmidt J, Assadi-Schmidt A, Bejinariu A, Müller P, Angendohr S, Babady M, Brinkmeyer C, Makimoto A, Kelm M. Performance of a convolutional neural network derived from an ECG database in recognizing myocardial infarction. *Sci Rep.* 2020;10(1):8445. <https://doi.org/10.1038/s41598-020-65105-x>.
13. Hae H, Kang SJ, Kim WJ, Choi SY, Lee JG, Bae Y, Cho H, Yang DH, Kang JW, Lim TH, Lee CH, Kang DY, Lee PH, Ahn JM, Park DW, Lee SW, Kim YH, Lee CW, Park SW, Park SJ. Machine learning assessment of myocardial ischemia using angiography: development and retrospective validation. *PLoS Med.* 2018;15(11):e1002693. <https://doi.org/10.1371/journal.pmed.1002693>.
14. Zreik M, Lessmann N, van Hamersvelt RW, Wolterink JM, Voskuil M, Viergever MA, Leiner T, Isgum I. Deep learning analysis of the myocardium in coronary CT angiography for identification of patients with functionally significant coronary artery stenosis. *Med Image Anal.* 2018;44:72–85. <https://doi.org/10.1016/j.media.2017.11.008>.
15. Hu X, Yang M, Han L, Du Y. Diagnostic performance of machine-learning-based computed fractional flow reserve (FFR) derived from coronary computed tomography angiography for the assessment of myocardial ischemia verified by invasive FFR. *Int J Cardiovasc Imaging.* 2018;34(12):1987–96. <https://doi.org/10.1007/s10554-018-1419-9>.
16. Du T, Xie L, Zhang H, Liu X, Wang X, Chen D, Xu Y, Sun Z, Zhou W, Song L, Guan C, Lansky AJ, Xu B. Automatic and multimodal analysis for coronary angiography: training and validation of a deep learning architecture. *EuroIntervention.* 2020. <https://doi.org/10.4244/EIJ-D-20-00570>.
17. Royer-Rivard R, Girard F, Dahdah N, Cheriet F. End-to-end deep learning model for cardiac cycle synchronization from multi-view angiographic sequences. *Annu Int Conf IEEE Eng Med Biol Soc.* 2020;2020: 1190–3. <https://doi.org/10.1109/EMBC41109.2020.9175453>.
18. Cook CM, Warisawa T, Howard JP, Keeble TR, Iglesias JF, Schampaert E, Bhindi R, Ambrosia A, Matsuo H, Nishina H, Kikuta Y, Shiono Y, Nakayama M, Doi S, Takai M, Goto S, Yakuta Y, Karube K, Akashi YJ, Clesham GJ, Kelly PA, Davies JR, Karamasis GV, Kawase Y, Robinson NM, Sharp ASP, Escaned J, Davies JE. Algorithmic versus expert human interpretation of instantaneous wave-free ratio coronary pressure-wire pull back data. *JACC Cardiovasc Interv.* 2019;12(14):1315–24. <https://doi.org/10.1016/j.jcin.2019.05.025>.
19. Priori SG, Blomström-Lundqvist C, Mazzanti A, Blom N, Borggrefe M, Camm J, Elliott PM, Fitzsimons D, Hatala R, Hindricks G, Kirchhof P, Kjeldsen K, Kuck KH, Hernandez-Madrid A, Nikolaou N, Norekval TM, Spaulding C, Van Veldhuizen DJ, ESC Scientific Document Group. ESC Guidelines for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death: the Task Force for the Management of Patients with Ventricular Arrhythmias and the Prevention of Sudden Cardiac Death of the European Society of Cardiology (ESC). Endorsed by: Association for European Paediatric and Congenital Cardiology (AEPC). *Eur Heart J.* 2015;36(41):2793–867. <https://doi.org/10.1093/eurheartj/ehv316>.
20. Motwani M, Dey D, Berman DS, Germano G, Achenbach S, Al-Mallah MH, Andreini D, Budoff MJ, Cademartiri F, Callister TQ, Chang HJ, Chinnaiyan K, Chow BJ, Cury RC, Delago A, Gomez M, Gransar H, Hadamitzky M, Hausleiter J, Hindoyan N, Feuchtner G, Kaufmann PA, Kim YJ, Leipsic J, Lin FY, Maffei E, Marques H, Pontone G, Raff G, Rubinstein R, Shaw LJ, Stehli J, Villines TC, Dunning A, Min JK, Slomka PJ. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. *Eur Heart J.* 2017;38(7):500–7. <https://doi.org/10.1093/eurheartj/ehw188>.
21. Zack CJ, Senecal C, Kinar Y, Metzger Y, Bar-Sinai Y, Widmer RJ, Lennon R, Singh M, Bell MR, Lerman A, Gulati R. Leveraging machine learning techniques to forecast patient prognosis after percutaneous coronary intervention. *JACC Cardiovasc Interv.* 2019;12(14):1304–11. <https://doi.org/10.1016/j.jcin.2019.02.035>.
22. Kwon JM, Jeon KH, Kim HM, Kim MJ, Lim S, Kim KH, Song PS, Park J, Choi RK, Oh BH. Deep-learning-based risk stratification for mortality of patients with acute myocardial infarction. *PLoS One.* 2019;14(10): e0224502. <https://doi.org/10.1371/journal.pone.0224502>.
23. Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JGF, Coats AJS, Falk V, González-Juanatey JR, Harjola VP, Jankowska EA, Jessup M, Linde C, Nihoyannopoulos P, Parisi JT, Pieske B, Riley JP, Rosano GMC, Ruilope LM, Ruschitzka F, Rutten FH, van der Meer P, ESC Scientific Document Group. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. *Eur Heart J.* 2016;37(27):2129–200. <https://doi.org/10.1093/eurheartj/ehw128>.
24. Gardin JM, Siscovich D, Anton-Culver H, Lynch JC, Smith VE, Klopstein HS, Bommer WJ, Fried L, O’Leary D, Manolio TA. Sex, age, and disease affect echocardiographic left ventricular mass and systolic function in the free-living elderly. *The Cardiovascular*

- Health Study. *Circulation*. 1995;91(6):1739–48. <https://doi.org/10.1161/01.cir.91.6.1739>.
25. McCullough PA, Philbin EF, Spertus JA, Kaatz S, Sandberg KR, Weaver WD, Resource Utilization Among Congestive Heart Failure (REACH) Study. Confirmation of a heart failure epidemic: findings from the Resource Utilization Among Congestive Heart Failure (REACH) study. *J Am Coll Cardiol*. 2002;39(1):60–9. [https://doi.org/10.1016/s0735-1097\(01\)01700-4](https://doi.org/10.1016/s0735-1097(01)01700-4).
 26. Ho KK, Pinsky JL, Kannel WB, Levy D. The epidemiology of heart failure: the Framingham study. *J Am Coll Cardiol*. 1993;22(4 Suppl A):6A–13A. [https://doi.org/10.1016/0735-1097\(93\)90455-a](https://doi.org/10.1016/0735-1097(93)90455-a).
 27. Solomon SD, Dobson J, Pocock S, Skali H, McMurray JJ, Granger CB, Yusuf S, Swedberg K, Young JB, Michelson EL, Pfeffer MA, Candesartan in Heart failure: Assessment of Reduction in Mortality and morbidity (CHARM) Investigators. Influence of nonfatal hospitalization for heart failure on subsequent mortality in patients with chronic heart failure. *Circulation*. 2007;116(13):1482–7. <https://doi.org/10.1161/CIRCULATIONAHA.107.696906>.
 28. Zhang J, Gajjala S, Agrawal P, Tison GH, Hallock LA, Beussink-Nelson L, Lassen MH, Fan E, Aras MA, Jordan C, Fleischmann KE, Melisko M, Qasim A, Shah SJ, Bajcsy R, Deo RC. Fully automated echocardiogram interpretation in clinical practice. *Circulation*. 2018;138(16):1623–35. <https://doi.org/10.1161/CIRCULATIONAHA.118.034338>.
 29. Tsao CW, Lyass A, Enserro D, Larson MG, Ho JE, Kizer JR, Gottdiner JS, Psaty BM, Vasan RS. Temporal trends in the incidence of and mortality associated with heart failure with preserved and reduced ejection fraction. *JACC Heart Fail*. 2018;6(8):678–85. <https://doi.org/10.1016/j.jchf.2018.03.006>.
 30. Shah SJ, Katz DH, Selvaraj S, Burke MA, Yancy CW, Gheorghiade M, Bonow RO, Huang CC, Deo RC. Phenomapping for novel classification of heart failure with preserved ejection fraction. *Circulation*. 2015;131(3):269–79. <https://doi.org/10.1161/CIRCULATIONAHA.114.010637>.
 31. Sengupta PP, Huang YM, Bansal M, Ashrafi A, Fisher M, Shameer K, Gall W, Dudley JT. Cognitive machine-learning algorithm for cardiac imaging: a pilot study for differentiating constrictive pericarditis from restrictive cardiomyopathy. *Circ Cardiovasc Imaging*. 2016;9(6):e004330. <https://doi.org/10.1161/CIRCIMAGING.115.004330>.
 32. Levy WC, Mozaffarian D, Linker DT, Sutradhar SC, Anker SD, Cropp AB, Anand I, Maggioni A, Burton P, Sullivan MD, Pitt B, Poole-Wilson PA, Mann DL, Packer M. The Seattle Heart Failure Model: prediction of survival in heart failure. *Circulation*. 2006;113(11):1424–33. <https://doi.org/10.1161/CIRCULATIONAHA.105.584102>.
 33. Panahiazar M, Taslimitehrani V, Pereira N, Pathak J. Using EHRs and machine learning for heart failure survival analysis. *Stud Health Technol Inform*. 2015;216:40–4.
 34. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. *J Am Med Inform Assoc*. 2017;24(2):361–70. <https://doi.org/10.1093/jamia/ocw112>.
 35. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, Pellikka PA, Enriquez-Sarano M, Noseworthy PA, Munger TM, Asirvatham SJ, Scott CG, Carter RE, Friedman PA. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat Med*. 2019;25(1):70–4. <https://doi.org/10.1038/s41591-018-0240-2>.
 36. Hernandez-Suarez DF, Kim Y, Villablanca P, Gupta T, Wiley J, Nieves-Rodriguez BG, Rodriguez-Maldonado J, Feliu Maldonado R, da Luz Sant'Ana I, Sanina C, Cox-Alomar P, Ramakrishna H, Lopez-Candales A, O'Neill WW, Pinto DS, Latib A, Roche-Lima A. Machine learning prediction models for in-hospital mortality after transcatheter aortic valve replacement. *JACC Cardiovasc Interv*. 2019;12(14):1328–38. <https://doi.org/10.1016/j.jcin.2019.06.013>.
 37. Zhu H, Cheng C, Yin H, Li X, Zuo P, Ding J, Lin F, Wnag J, Zhou B, Li Y, Hu S, Xiong Y, Wang B, Wan G, Yang X, Yuan Y. Automatic multilabel electrocardiogram diagnosis of heart rhythm or conduction abnormalities with deep learning: a cohort study. *Lancet Digit Health*. 2020;2:e348–57.
 38. Ribeiro AH, Ribeiro MH, Paixão GMM, Oliveira DM, Gomes PR, Canazart JA, Ferreira MPS, Andersson CR, Macfarlane PW, Meira W Jr, Schön TB, Ribeiro ALP. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat Commun*. 2020;11(1):1760. <https://doi.org/10.1038/s41467-020-15432-4>. Erratum in: *Nat Commun*. 2020 May 1;11(1):2227.
 39. Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, Carter RE, Yao X, Rabinstein AA, Erickson BJ, Kapa S, Friedman PA. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet*. 2019;394(10201):861–7. [https://doi.org/10.1016/S0140-6736\(19\)31721-0](https://doi.org/10.1016/S0140-6736(19)31721-0).
 40. Anuradha MB, Reddy VCV. Cardiac arrhythmia classification using fuzzy classifiers. *J Theor Appl Inform Tech*. 2009;4:353–9.
 41. Alfaras M, Soriano MC, Ortín S. A fast machine learning model for ECG-based heartbeat classification and arrhythmia detection. *Front Phys*. 2019;7:103.
 42. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, Ng AY. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med*. 2019;25(1):65–9. <https://doi.org/10.1038/s41591-018-0268-3>. Epub 7 Jan 2019. Erratum in: *Nat Med*. 2019 Mar;25(3):530.
 43. Eerikäinen LM, Vanschoren J, Rooijakkers MJ, Vullings R, Aarts RM. Reduction of false arrhythmia alarms using signal selection and machine learning.

- Physiol Meas. 2016;37(8):1204–16. <https://doi.org/10.1088/0967-3334/37/8/1204>.
44. Krasteva V, Jekova I, Leber R, Schmid R, Abächerli R. Real-time arrhythmia detection with supplementary ECG quality and pulse wave monitoring for the reduction of false alarms in ICUs. Physiol Meas. 2016;37(8):1273–97. <https://doi.org/10.1088/0967-3334/37/8/1273>.
45. Heaven D. Why deep-learning AIs are so easy to fool. Nature. 2019;574(7777):163–6. <https://doi.org/10.1038/d41586-019-03013-5>.



Sara Moccia and Elena De Momi

Contents

Introduction	826
Pre-operative Planning	827
Pre-/Intra-operative Registration	828
Execution	829
Intra-operative Image Analysis	830
Monitoring and Assessment	830
Conclusion	832
Cross-References	832
References	832

Abstract

Medical robotics emerged in the 1980s to improve clinicians' technical capability and increase safety in clinical procedures. This chapter specifically addresses the topic of surgical robots. Major opportunities for artificial

intelligence (AI) here include (i) surgical planning, (ii) intra-operative registration, (iii) surgical execution, and (iv) surgery evaluation/assessment. This chapter provides an overview of AI methodologies developed so far in these four fields, reporting main limitations and open challenges. Deep learning (DL) models for pre-operative image segmentation, classification, and detection are presented, along with DL architectures for intra-operative registration. In the context of surgical execution, intra-operative image analysis is described, focusing on endoscopic images for tissue and surgical tool segmentation. Considering the rapid evolution of AI applications in the field of surgical robotics, the proposed contribution is aimed at giving to young researchers and surgeons working in the field of surgical

S. Moccia (✉)

The BioRobotics Institute, Scuola Superiore Sant'Anna,
Pisa, Italy

Department of Excellence in Robotics & AI, Scuola
Superiore Sant'Anna, Pisa, Italy
e-mail: sara.moccia@santannapisa.it

E. De Momi

Department of Electronics, Information and
Bioengineering, Politecnico di Milano, Milan, Italy
e-mail: elenade.momi@polimi.it

robotics a complete overview on the current AI applications proposed in literature.

Keywords

Medical robotics · Artificial intelligence · Surgical planning · Intra-operative registration · Intra-operative image analysis · Virtual-fixture control

Introduction

Medical robotics is an interdisciplinary field that emerged in 1980 as a new branch of robotics. A large number of challenges has to be tackled to bring robots into the clinical routine, including high accuracy for surgical procedures and the need for human-robot interfaces. These challenges are far from those of industrial robots, and this motivated the birth of the new field [1].

Nowadays, robots are used in several domains, including neurosurgery, orthopedic surgery, nose and throat surgery, and laparoscopy. Medical robots allow improving surgeons' technical capability to perform procedures by exploiting the complementary strengths of humans and robots: robots can be more precise and geometrically accurate than unaided average surgeons. Medical robots can also increase surgical safety by means

of forbidden-region, virtual-fixture control (e.g., to avoid surgical tools damaging sensitive structures). As a last advantage, medical robots allow capturing detailed procedural data, which can be analyzed to update/adjust the current surgical practices to improve the surgical outcome.

Broadly, surgical robots may be grouped into:

- *Surgeon extender robots*, which manipulate surgical instruments under the direct control of the surgeon, usually through a teleoperated or hands-on cooperative control interface, and allow tremor filtering, region-avoidance/targeting control and remote surgery. Examples include the da Vinci system (Intuitive Surgical Systems, Sunnyvale, CA) and the Steady Hand microsurgery robot (a research prototype from Johns Hopkins University).
- *Auxiliary surgical support robots*, which generally work alongside the surgeon and perform routine tasks, such as endoscope holding. Examples include the AESOP endoscope positioner.

Following the nomenclature proposed in [2], the main aspects of surgical robots are (Fig. 1) surgical planning (section “[Pre-operative Planning](#)”), registration (section “[Pre-/Intra-operative Registration](#)”), execution (section “[Execution](#)”), and evaluation (section “[Monitoring and Assessment](#)”), which all

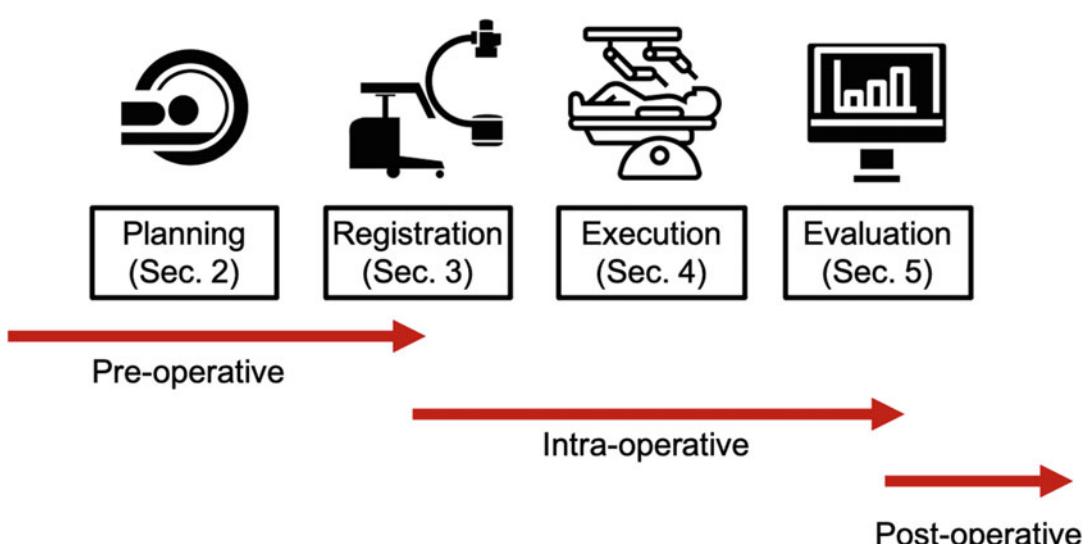


Fig. 1 Standard workflow of robotic surgery

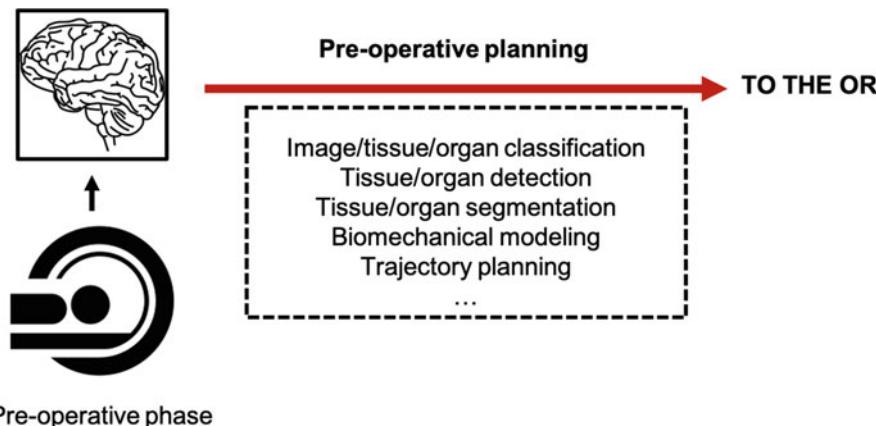


Fig. 2 Pre-operative surgical planning. OR: operating room

together can be referred to as CAD (computer-aided design)/CAM (computer-aided manufacturing). This chapter will cover all these aspects, focusing on specific points in which artificial intelligence (AI) technologies mainly contribute.

Pre-operative Planning

Surgical planning relies on pre-operative imaging, such as magnetic resonance imaging (MRI) and computerized tomography (CT), to provide anatomical/morphological information on the patient (Fig. 2). Surgical planning covers a major role for percutaneous keyhole surgeries (e.g., liver-tumor ablation) and stereotactic procedures (e.g., key-hole neurosurgery). Such surgeries require accurate positioning of the instruments to guarantee an effective and safe procedure, but usually do not provide a clear view of the surgical scenario. Surgical planning is also crucial for orthopedic and cardiac surgery, where it allows choosing the best prosthesis size and its best positioning with respect to the patient's anatomy. In laparoscopy, surgical planning allows estimating the optimal location of the skin incision, as well as the optimal placement of surgical instruments.

Different researchers are working on developing surgical planners for helping surgeons in processing pre-operative images and performing surgical planning [3]. Pre-operative planning involves a variety of aspects, ranging from biomechanical modeling to

tissue segmentation, classification, and detection. While AI for biomechanical modeling is at its infancy, AI covers today a major and undisputed role in image segmentation, classification, and detection [4]. Classification involves images (e.g., to understand if a structure of interest is present in the image) and object/lesion classification (e.g., for staging tumors). Standard deep learning (DL) architectures, mostly convolutional neural networks (CNNs) pre-trained on large natural datasets such as ImageNet, are commonly used. Examples include VGG16, AlexNet, ResNet, and Google Inception [5].

Opposite to classification, detection spatially locates objects in the image. The problem of detection is called localization if only one object has to be found in the image. Main applications include tissue, organ, and surgical tool detection. Here, region-proposal architectures, such as RCNN, Fast RCNN, and Faster RCNN [6], as well as detection architectures inspired by Yolo [7], are the most exploited.

Image segmentation is mostly applied for segmenting organs and pathological tissues (e.g., cancerous areas). A plethora of segmentation algorithms today exist. The U-Net [8] is among the first CNN models for medical image segmentation. With U-Net, the concept of encoder-decoder architectures with long skip connections has been introduced to allow accurate and fast segmentation. Modern algorithms involve more complex architectures, following the adversarial training paradigm [9]. Besides

tissue segmentation, AI algorithms can be also used to estimate morphological and metabolic parameters, such as tissue oxygenation [10], starting from metabolic images such as positron emission tomography (PET).

AI in the field of pre-operative planning finds space also in trajectory optimization for keyhole surgeries. Examples include the application of machine learning to predict the entry and target regions [11], as well as the optimal trajectory to minimize the probe depth while maximizing the distance from critical structures [12]. A systematic review of surgical planning assistance in keyhole and percutaneous surgery has been recently published [3]. For brain surgery, large attention is given to the localization of stimulation zones, to estimate the size and location of tissue that has to be removed or treated [13]. This can be done by exploiting additional signals, such as electroencephalography and functional near-infrared spectroscopy. A systematic review on the possibilities offered by AI in brain intervention planning (i.e., from diagnosis to intervention) has been published [12].

Pre-/Intra-operative Registration

In the operating room, the pre-operative patient model and plan must be registered to the actual patient. With registration, the pre-operative

patient anatomy is aligned with the intra-operative view. The workflow of image registration is shown in Fig. 3. The registration process can be described by a mathematical operator called transformation, which can be rigid or deformable. Rigid transformation is described by 6 degrees of freedom (3 for rotation and 3 for translation), and the Iterative Closest Point is the most used rigid transformation algorithm. Deformable registration can be described up to infinite degrees of freedom [14].

There is an extensive literature on techniques for coregistering coordinate systems associated with robots, sensors, images, and the patient. Typically, the transformation is estimated by identifying corresponding landmarks or structures on the pre-operative model and the patient by using (i) additional imaging, (ii) tracked devices, and (iii) the robot itself [2]. If the patient's anatomy has changed from the pre-operative plan, then the planning needs to be updated.

The role of AI in registration has impacted especially the field of image-based registration, where DL has changed the landscape of image registration research. Today, more and more advanced imaging devices are introduced in the operating room [15]. Such devices produce images that may be quite different from those acquired pre-operatively. Typically, intra-operative imaging modalities are X-ray,

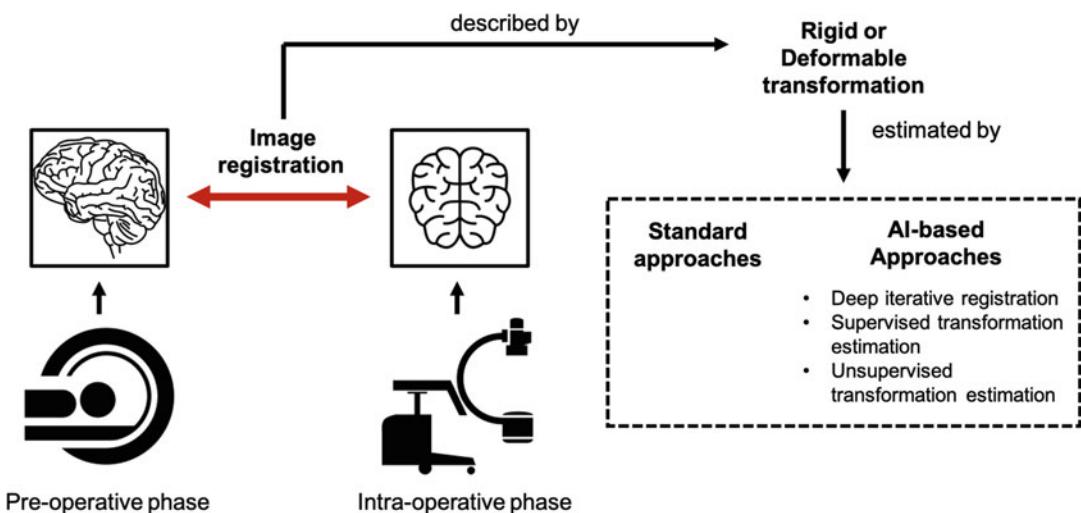


Fig. 3 Pre-operative to intra-operative image registration

ultrasound, and video laparoscopy. This poses additional issues for the registration process, as multi-modal registration has to be performed.

First approaches dealing with AI and registration were used to improve the performance of standard intensity-based registration algorithms. The need to shorten the computational time of registration algorithms motivated researchers to move forward, by designing one-step transformation estimation techniques, mostly with unsupervised methods to tackle the lack of generating ground truth data.

Registration using AI can be categorized in:

- *Deep iterative registration*, among the first approaches in the field, where researchers use DL to learn a similarity metric between images. This similarity metric is used in a classical intensity-based registration framework.
- *Supervised transformation estimation*, where a supervised DL model is trained to estimate in one step the transformation parameters:
 - Fully Supervised Transformation Estimation, where the DL model is fully supervised. In this case, the quality of the registrations is clearly dependent on the availability and quality of the ground truth. Usually, synthetic data are used to overcome such limitations.
 - Dual/Weakly Supervised Transformation Estimation, where the dual supervision refers to using both ground truth data and some image similarity metrics to train the DL model, while weak supervision refers to using the segmentation of corresponding regions in the pre- and intra-operative images for designing the DL model loss function. This approach partially attenuates the limitation of methods for fully supervised transformation estimation, representing one of the most promising methods in the field of image registration.
- *Unsupervised transformation estimation*, where the transformation is estimated without the need of an annotated ground truth. The power of unsupervised estimation has been unlocked by the implementation of the spatial transformer network, which explicitly allows

the spatial manipulation of images (and activation maps) within the DL network [16]:

- Similarity Metric-Based Unsupervised Transformation Estimation, where similarity metrics are used to train the DL network. Approaches in the field make use of both standard similarity metrics (e.g., normalized cross-correlation) and more advanced ones (e.g., based on the generator output of a generative adversarial network). Unfortunately, today the problem of quantifying image similarity for multimodal registration applications is still open. This poses issues considering that pre-operative to intra-operative image registration is often a multimodal registration problem.
- Feature-Based Unsupervised Transformation Estimation, where the feature representations of the images are used to train the DL network. This approach is significantly more difficult than the previous ones, as the estimation of the registration has to be performed starting from the estimation of feature representations.

Execution

As introduced in section “[Pre-/Intra-operative Registration](#),” modern robotic surgical rooms are equipped with multimodal imaging systems, to provide all possible support to the surgical team. Intra-operative ultrasounds, X-ray, optical coherence tomography, interventional MR or X-ray imaging, and endoscopy are among the most exploited imaging modalities and are rapidly evolving to provide more and more accurate information. New imaging modalities, such as diffuse reflectance spectroscopy and multispectral imaging, will enter the actual surgical practice in the next decades. Each modality has its own characteristics in terms of field of view (FoV), image resolution, noise, and enhanced tissues. Besides imaging systems, the operating room is equipped with a variety of other medical devices, which provide quantitative and continuous information on the patient status. The robot itself provides information on the surgical procedure.

The centrality of AI for intra-operative data analysis is undeniable. During the last decade, a new discipline has emerged, called surgical data science (SDS), which exploits AI for processing intra-operative data analysis, as to provide surgeons with decision support and context-aware assistance [17] in robotic surgery. This section describes some of the major opportunities for SDS in intra-operative image analysis (section “[Intra-operative Image Analysis](#)”), with a focus on endoscopy.

Intra-operative Image Analysis

Intra-operative image analysis for endoscopic images may be classified in the following categories:

- Informative frame selection
- Detection, classification, and segmentation of anatomical structures and surgical tools

Such categories will be shortly described hereafter, while a more in-depth review on the topic can be found in [18].

Informative frame selection. From the surgeon’s side, reviewing an endoscopic video is a focus-intensive operation. While focusing on particular structures during the video examination, clinicians may miss important clues indicating suspicious conditions (e.g., early tissue alterations). This process could be further compromised by the presence of uninformative video portions. Developing a strategy to select informative frames has the potential to reduce the amount of data to review, lowering the surgeons’ workload. Several approaches have been proposed to select informative video frames, from (time-consuming) manual frame selection to random or uniform frame selection, which is fast but does not guarantee that all informative frames are extracted. More recently, DL approaches have shown promising results.

Detection, classification, and segmentation of anatomical structures and surgical tools. The analysis of an intra-operative informative frame usually starts with the *classification, detection,*

and segmentation of structures in the FoV. This means to understand if a specific structure (such as a pathological tissue or a surgical tool) is present in an image. Structure detection and classification have strong diagnostic value but can be also used to determine the surgical phase (hence, in each phase of the surgery, different structures are present in the FoV). The most used approaches in the field are similar to those described in section “[Pre-operative Planning](#).” The main difference here is that endoscopic videos naturally encode the temporal information. Hence, models to analyze this information are becoming more and more studied by the researchers. Examples include recurrent CNNs and spatiotemporal CNNs [19]. Here major challenges are relevant to the higher complexity of models analyzing temporal clips (with several frames), opposite to those processing still frames. This may pose issues relevant to overfit. To attenuate this issue, short temporal clips (up to 10 frames) are used. In the field of intra-operative segmentation, interesting results are achieved by adversarial models [20, 21].

Providing automatic interpretation of the surgical scene would be fundamental to increase the level of autonomy of robots in surgery [22]. The approach presented in [23] combines a deductive approach for surgical step classification with an inductive approach derived from an ontological description of the surgical process (i.e., partial nephrectomy). Recently, some groups have started integrating logic formalism and rules management into the surgical process in order to allow for continuous learning of new rules by performing the surgical tasks [24].

Monitoring and Assessment

The role of AI in monitoring and assessing surgical procedures is growing rapidly. Surgical skill assessment is among the major opportunities in the field. Surgical skill assessment is particularly useful during surgeon training [25]. Nowadays most assessments of trainee skills are still

performed via outcome-based analysis or structured checklists. Modern surgical robots are able to collect a large amount of sensory data: such data represent a unique opportunity for evaluating the skills and proficiencies of the trainees.

Surgeon's skill levels can be defined according to the Dreyfus model [26], shown in Table 1. The role of AI algorithms for surgical skill assessment is to abstract the Dreyfus model by analyzing data available in the operating room.

To design algorithms for automatic surgical skill assessment, researchers usually take inspiration from available standards used in clinics, such as the Objective Structured Assessment of Technical Skills (OSATS). The evaluation here relies on respect for tissue (used forces, caused damage), time and motion (efficiency), instrument handling (movement fluidity), knowledge of instruments (types and names), flow of operation (stops frequency), use of assistants (proper

strategy), and knowledge of specific procedure (familiarity with the aspect of the procedure). Such aspects can be derived from (Fig. 4):

- Kinematic data (e.g., instrument travel time, path length, velocity)
- Endoscopic videos
- Additional sensors (e.g., force sensors)

When using specific robots, more data may be available. Examples include system events, such as frequency of master controller clutch use, third-arm swap, and energy use, for the da Vinci system.

Once these data are available, algorithms to process them have to be designed. Approaches in the literature can be mainly divided in [27]:

- Descriptive statistic analysis, which aims to compute relevant indicators from surgical data to quantitatively describe skill level via

Table 1 Dreyfus model of skill acquisition. Columns and rows refer to mental function and skill level, respectively

	Novice	Competent	Proficient	Expert	Master
Recollection	<i>Non-situational</i>	<i>Situational</i>	<i>Situational</i>	<i>Situational</i>	<i>Situational</i>
Recognition	<i>Decomposed</i>	<i>Decomposed</i>	<i>Holistic</i>	<i>Holistic</i>	<i>Holistic</i>
Decision	<i>Analytical</i>	<i>Analytical</i>	<i>Analytical</i>	<i>Intuitive</i>	<i>Intuitive</i>
Awareness	<i>Monitoring</i>	<i>Monitoring</i>	<i>Monitoring</i>	<i>Monitoring</i>	<i>Absorbed</i>

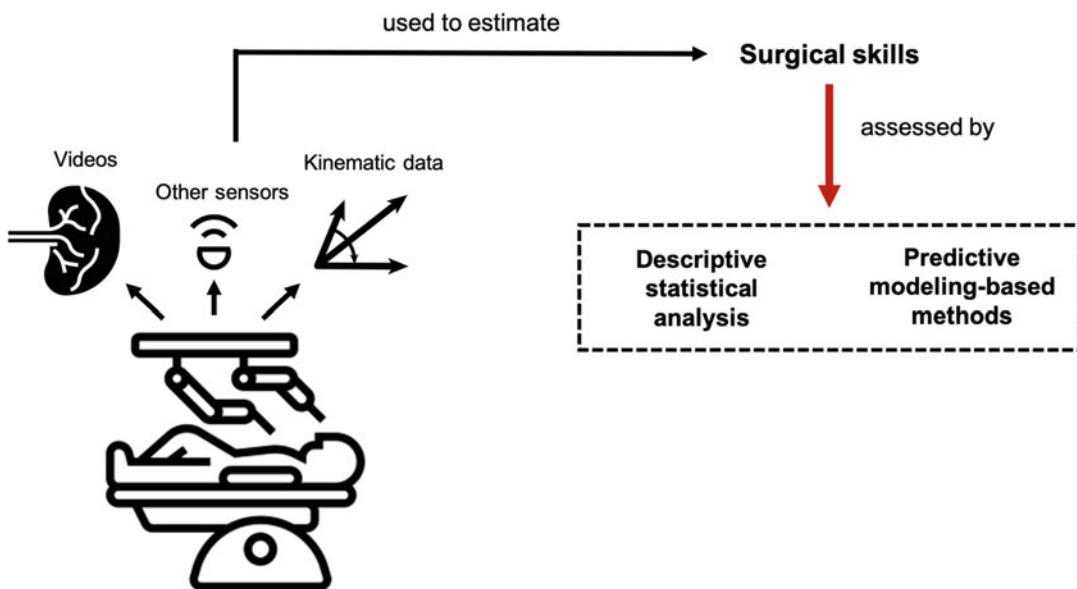


Fig. 4 Surgical skill assessment workflow

- statistical analysis. Sample indicators are path length, motion jerk, and tool orientation. However, selecting the most suited indicators may not be trivial considering the intra- and inter-variability in surgical procedure execution.
- Predictive modeling-based methods, which work on raw data to automatically extract and classify surgeon skill level. Here, AI covers a major role.

Among predictive modeling-based methods, a class of algorithms starts from features extracted following the OSATS guidelines and applies traditional machine-learning classifiers (e.g., support vector machines, random forest, k-nearest neighbors) to classify the skill level. To take the temporal information into account, hidden Markov models are commonly exploited. Focusing on video data, similarly to what introduced in section “[Intra-operative Image Analysis](#),” DL models processing 2D + t data (e.g., spatiotemporal CNNs or recurrent networks) are the most used approaches. DL models are commonly used to classify short video sequences (to avoid training large models, which are prone to overfitting) and are then followed by a consensus layer to predict the overall surgical procedure skill level.

Conclusion

In this chapter, the main steps of robotic surgery have been presented. The field of surgical robotics is evolving rapidly, thanks to the cooperation of surgeons, research institutions, and industry deploying surgical robots, with the final goal to have a direct impact on patients’ health.

In the close future, surgical robots need to become more precise, dexterous, and sensitive while also becoming much more compact and inexpensive. At the same time, surgical robots should be highly modular architectures that allow a high degree of reuse, with open interfaces between major subsystems. These are two main challenges that need to be tackled from the hardware perspective.

The use of AI for surgical robotics probably sees its main application in imaging, modeling,

and analysis: AI can help building patient-specific anatomical models from pre-operative images and real-time sensor data, for incorporating biomechanical information into these models and for using this information to help control the robot.

Cross-References

- [AIM in Endoscopy Procedures](#)

References

- Schweikard A, Ernst F. Medical robotics. Heidelberg: Springer; 2015.
- Taylor RH, Menciassi A, Fichtinger G, Fiorini P, Dario P. Medical robotics and computer-integrated surgery. In :Springer handbook of robotics. Cham: Springer; 2016. p. 1657–84.
- Scorza D, El Hadji S, Cortés C, Bertelsen Á, Cardinale F, Baselli G, … De Momi E. Surgical planning assistance in keyhole and percutaneous surgery: a systematic review. *Med Image Anal.* 2020. 101820.
- Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, … Sánchez CI. A survey on deep learning in medical image analysis. *Med Image Anal.* 2017;42:60–88.
- Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Commun ACM.* 2017;60(6):84–90.
- Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2016. 39(6):1137–1149.
- Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 779–88.
- Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Cham: Springer; 2015. p. 234–41
- Xue Y, Xu T, Zhang H, Long LR, Huang X, Segan: adversarial network with multi-scale l1 loss for medical image segmentation. *Neuroinformatics.* 2018;16(3–4):383–92.
- Fan AP, An H, Moradi F, Rosenberg J, Ishii Y, Nariai T, … Zaharchuk G Quantification of brain oxygen extraction and metabolism with [15O]-gas PET: a technical review in the era of PET/MRI. *NeuroImage.* 2020. 117136.
- Li K, Vakharia VN, Sparks R, França LG, Granados A, McEvoy AW, … Duncan JS. Optimizing trajectories for cranial laser interstitial thermal therapy using

- computer-assisted planning: a machine learning approach. *Neurotherapeutics*. 2019;16(1):182–91.
- 12. Segato A, Marzullo A, Calimeri F, De Momi E. Artificial intelligence for brain diseases: a systematic review. *APL Bioeng*. 2020;4(4):041503.
 - 13. Favaro A., Segato A., Muretti F., De Momi E. An evolutionary-optimized surgical path planner for a programmable bevel-tip needle. *IEEE Trans Robot*. 2021.
 - 14. Haskins G, Kruger U, Yan P. Deep learning in medical image registration: a survey. *Mach Vis Appl*. 2020;31(1):8.
 - 15. Zaffino P, Moccia S, De Momi E, Spadea MF. A review on advances in intra-operative imaging for surgery and therapy: imagining the operating room of the future. *Ann Biomed Eng*. 2020;1–21.
 - 16. Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks. In: Advances in neural information processing systems. APA 2015. p. 2017–25.
 - 17. Maier-Hein L, Vedula SS, Speidel S, Navab N, Kikinis R, Park A, ... Hashizume M. Surgical data science for next-generation interventions. *Nat Biomed Eng*. 2017;1(9):691–6.
 - 18. Moccia S, Romeo L, Migliorelli L, Frontoni E, Zingaretti P. Supervised CNN strategies for optical image segmentation and classification in interventional medicine. In: Deep learners and deep learner descriptors for medical applications. Cham: Springer; 2020. p. 213–36.
 - 19. Colleoni E, Moccia S, Du X, De Momi E, Stoyanov D. Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers. *IEEE Robot Autom Lett*. 2019;4(3):2714–21.
 - 20. Casella A, Moccia S, Frontoni E, Paladini D, De Momi E, Mattos LS. Inter-foetus membrane segmentation for TTTS using adversarial networks. *Ann Biomed Eng*. 2020;48(2):848–59.
 - 21. Marzullo A, Moccia S, Catellani M, Calimeri F, De Momi E. Towards realistic laparoscopic image generation using image-domain translation. *Comput Methods Prog Biomed*. 2020;105834.
 - 22. Attanasio A, Scaglioni B, De Momi E, Fiorini P, Valdastri P. Autonomy in surgical robotics. *Annu Rev Control Robot Auton Syst*. 2020;4:651–679.
 - 23. Nakawala H, Bianchi R, Pescatori LE, De Cobelli O, Ferrigno G, De Momi E. “Deep-onto” network for surgical workflow and context recognition. *Int J Comput Assist Radiol Surg*. 2019;14(4):685–96.
 - 24. Meli D, Fiorini P, Sridharan M. Towards inductive learning of surgical task knowledge: a preliminary case study of the peg transfer task. *Procedia Comput Sci*. 2020;176:440–9.
 - 25. Mariani A, Pellegrini E, De Momi E. Skill-oriented and performance-driven adaptive curricula for training in robot-assisted surgery using simulators: a feasibility study. *IEEE Trans Biomed Eng*. 2021;68(2):685–694. <https://doi.org/10.1109/TBME.2020.3011867>.
 - 26. Nagyné Elek R, Haidegger T. Robot-assisted minimally invasive surgical skill assessment – manual and automated platforms. *Acta Polytech Hungarica*. 2019;16(8):141–69.
 - 27. Wang Z, Fey AM. Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. *Int J Comput Assist Radiol Surg*. 2018;13(12):1959–70.



AI in Surgical Robotics

60

Samyakh Tukra, Niklas Lidströmer, Hutan Ashrafian, and
Stamatia Giannarou

Contents

Introduction	836
Cognitive Surgical Robots	837
Proprioception	837
Depth Perception	837
Navigation	840
Surgical Tool Tracking	843
Haptic Feedback and Tissue Interaction Sensing	843
Advanced Visualization with Augmented Reality	844

S. Tukra (✉)
Department of Surgery and Cancer, Imperial College
London, London, UK
e-mail: samyakh.tukra17@imperial.ac.uk;
samtukra@thirdeye.health

N. Lidströmer
Department of Women's and Children's Health, Karolinska
Institutet, Stockholm, Sweden
e-mail: niklas.lidstromer@ki.se; niklas@lidstromer.com

H. Ashrafian
Department of Surgery and Cancer, Imperial College
London NHS Trust, London, UK

Institute of Global Health Innovation, Imperial College
London, London, UK

Hamlyn Centre for Robotics and Artificial Intelligence,
Department of Surgery and Cancer, Imperial College
London, London, UK
e-mail: h.ashrafian@imperial.ac.uk

S. Giannarou
Hamlyn Centre for Robotic Surgery and AI, Department of
Surgery and Cancer, Imperial College London, London,
UK
e-mail: stamatia.giannarou@imperial.ac.uk

Robot-Assisted Task Execution	845
Context-Aware Decision Support	846
Outlook	849
References	849

Abstract

The future of surgery is tightly knit with the evolution of artificial intelligence (AI) and its thorough involvement in surgical robotics. Robotics long ago became an integral part of the manufacturing industry. The area of healthcare though adds several more layers of complication. In this chapter we elaborate a broad range of issues to be dealt with when a robotic system enters the surgical theater and interacts with human surgeons – from overcoming the limitations of minimally invasive surgery to the enhancement of performance in open surgery. We present the latest from the fields of cognitive surgical robots, focusing on proprioception, intraoperative decision-making, and, ultimately, autonomy. More specifically, we discuss how AI has advanced the research field of surgical tool tracking, haptic feedback and tissue interaction sensing, advanced intraoperative visualization, robot-assisted task execution, and finally land in the crucial development of context-aware decision support.

Introduction

Robots are a crucial part of multiple industries, allowing greater productivity and efficiency by taking over repetitive, labor-intensive, menial, high-power, and dangerous tasks. This field of complex control systems has had a huge torrent in adoption to surgery, due to the key benefits it poses such as less pain, scarring, risk of infection, blood loss, and high precision, especially when operating in confined spaces or in minute settings like ophthalmic surgery. However, in reality, most robotic systems that are employed today are primarily machines with manipulators that are pre-programmed for conducting a set of predefined manual tasks automatically. This also requires prior knowledge of the environment so that the

robot will be deployed in such a way that all possible changes and interactions can be modeled and defined in advance. This became possible with the advent of sensors embedded in the robots that provide information about the environment and feedback of performance in accomplishing the given task.

However, robotic autonomy in a highly dynamic environment like surgery where external manipulations that occur in the form of interactions with the patient is challenging. This is because soft tissue deformation, varying material properties, and the manipulation of anatomical structures with respect to the robotic tool cannot be pre-programmed or defined as a set of instructions over time. Hence, the prior motivation of robotic development that was mainly to increase productivity and reduce costs by tackling menial, repetitive tasks needs to be challenged. This is exactly what the field of intelligent systems or in particular Artificial Intelligence (AI) aims to achieve. Research in AI has enabled endowing robots with the ability to adapt to a changing environment by analyzing their sensor data to have greater perception and adaptability to an uncertain environment. Additionally, AI enables them to recognize hidden patterns in data at high dimensions, such that they can autonomously determine the optimal action even in cases difficult for humans.

Evidently, AI is essential for the evolution of robotics to cognitive robotics. Here, AI is the information processing engine like the brain, and robotics gives it a physical form to interact with the environment. Despite the high synergy between robotics and AI, yet the two fields progressed distinctly apart in the previous few years. Recent advances in computer vision, machine and deep learning, data analytics, and robotics have enabled the development of cognitive robotic systems capable of operating at varying degrees of autonomy to assist the surgeon intraoperatively and improve the efficacy and outcome of surgery.

In this chapter, the application of AI in the field of surgical robotics will be extensively discussed. In particular, we will explain how we can use AI to endow robots with cognitive abilities such as proprioception and decision-making and how these methodologies pave the way for robotic autonomy to enhance the capabilities of the surgeon and improve surgical outcome. The following sections discuss the enabling technologies of AI to achieve the above goal focusing on navigation and 3D scene understanding, advanced visualization using augmented reality, followed by diagnosis support for intraoperative decision-making. Finally, we present an outlook into the future.

Cognitive Surgical Robots

For robots to be capable of interacting with the surgical scene in a safe manner, it is vital we embed them with intelligence and perception power similar to that of humans. As an example, humans have exemplar ability of proprioception and thereby make decisions on self-motion in a dynamic environment. This is achieved through the processing of many neural impulses the brain receives via the sensory inputs, thereby making the human aware of the local surrounding and make decisions on how to navigate themselves in complex three-dimensional space with ease. Providing robots with the necessary cognitive capabilities to perceive their environment and feedback that information will enable decision-making and ultimately robot autonomy.

The degree of autonomy achievable by a surgical robot ranges from robot assistance (Level 1) to conditional autonomy (Level 3) and full autonomy (Level 5) at the highest level [1]. Each of these autonomy levels is characterized by different technical challenges and therefore relies on different technologies for successful execution. Surgical platforms operating at low degree of autonomy (Level 1) provide the surgeon with either cognitive or physical assistance without taking control of the action being performed. Examples include systems that assist in optimal robot deployment and systems that provide augmented reality. To achieve this goal, the robot needs to have a good understanding

of the surgical scene. More specifically, the 3D structure of the surgical scene, the tracking and recognition of surgical tools as well as their interaction with the tissue are fundamental for intraoperative robotic assistance. For seamless augmented reality visualization, the detection of anatomical landmarks and the registration of multimodal data are required.

Task autonomy (Level 2) enables a surgical robot to execute repetitive and well-defined tasks that may be ergonomically difficult for the surgeon, but without being able to update the task planning during execution. Robot-assisted execution of surgical tasks such as tissue retraction, suturing, and ablation require the analysis of the surgical workflow to permit the robot to provide dedicated support at the right phase of the operation.

As the degree of autonomy increases, surgical robots operate with conditional autonomy (Level 3), being capable of planning how to execute a surgical task and updating the plan during execution. To achieve this, the robot is constantly monitoring the surgical environment to extract contextual information and execute the task in real time. Key AI technologies to enable this involve modeling of the tissue deformation and high-level feature tracking.

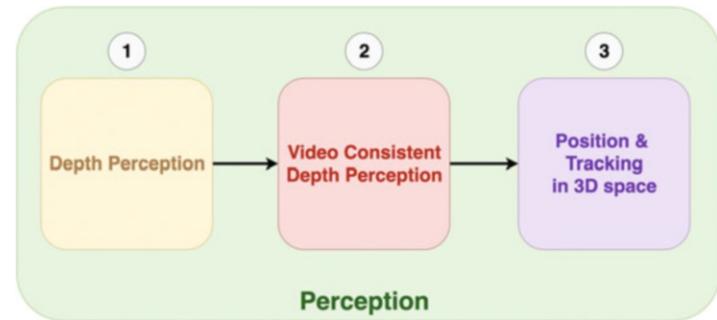
The robotic platforms operating at high autonomy (Level 4), have the ability to make clinical decisions and execute them under human supervision. As a future application, they could be used for the intelligent removal of cancerous tissue, including identification of the tissue state, adaptation of the plan with real-time data, and ablation of the cancer while sparing as much healthy tissue as possible. To this end, AI techniques for tissue characterization have a crucial role toward tumor detection and resection. Finally, fully autonomous surgical robotic platforms (Level 5) have not been developed yet.

Proprioception

Depth Perception

This area of research relies heavily on computer vision and in particular extracting 3D models of the scene. Three-dimensional information is

Fig. 1 Process pipeline of embedding proprioception in robotics from computer vision



quintessential for registering preoperative data to the surgical field of view, which can then be further visualized through mixed reality. These steps are broken down and illustrated in Fig. 1.

The core aim in the first instance is to generate a dense 3D model of the scene captured through a camera. To get a 3D model of the surgical scene we essentially need to calculate the distance of various points in the image relative to the position of the camera. This not only provides the information captured in the image but also its 3D location in the scene. This information is referred to as a “depth map,” a single channel matrix that encapsulates the distance value for each respective pixel in the image. The general method of depth information extraction from images is to acquire a pair of images using two cameras displaced from each other by a known distance called the “baseline.” This pair of images is known as stereo-images, which are similar to that of a biological visual system, where the eyes together are capable of establishing 3D information of what you see. The main idea is as follows: by estimating the per-pixel displacement between the two frames horizontally, commonly known as the “disparity map,” depth can be inferred as an inverse function. As an alternative, two or more images taken from a moving camera can also be used to compute depth information. This method of dealing with a moving camera and multiple images is known as Structure from Motion (SfM), discussed further in the next subsection 3.2. Hence, the methods of extracting depth from an image can be divided into the following categories: single image depth (mono-depth), stereo image depth (binocular-depth), and multiple views (SfM). Recently, deep learning models

have become very attractive in the depth perception workflow, providing end-to-end solutions for depth estimation.

Single image depth: Traditional vision uses multiple frames for inferring depth and thus, one would state extracting depth from a single image is an ill-posed problem. However, if humans were to cover one of their eyes, they are still capable of perceiving depth even from a single vision system, as they are capable of using priori information of their surroundings. The seven monocular depth cues are: relative size, interposition, linear perspective, aerial perspective, light and shade, and monocular movement parallax. Hence, this argument motivates a lot of the work in single image depth estimation. Approaches for depth estimation fall into two categories, namely, supervised learning based, which require ground truth depth data and self-supervised learning based, which do not rely on any ground truth for training.

Supervised learning based approaches, frame depth estimation either as a classification or a regression problem, since ground truth labels are available. The loss signal is typically some type of distance metric such as L1, L2, etc., between the predicted depth map and the ground truth depth map. One of the earliest works predicted dense depth maps via a 2-stage process, where in the first stage a deep neural network produces a coarse depth map followed by the second stage where another deep neural network refines it [2].

In the first stage the coarse depth map is based on the network learning global features from images, like object locations, coarse outlines, etc. In the second stage the finer scales are estimated by the network learning more local features such as textures of specific objects, edges of local

points, etc. The authors treated the problem as a regression task and defined losses that were invariant to image scale. Other methods were later published that built upon Eigen et al. [2], such as Cao and Shen [3], which treated the problem as a classification task of the valid pixels in the ground truth depth maps. The authors of [4] deviated from the aforementioned approach and treated monocular depth estimation as an ordinal regression problem. Modeling traditional regression onto an ordinal scale means any predicted variable can be from any arbitrary scale. Hence, only the ordering between intermediate values is vital to capture the relationship of the variable. This enables the model to discretize a continuous depth map into intervals, which is beneficial since standard regression-based losses result in models suffering from slow convergence or convergence to local extrema. However, supervised approaches depend upon high quality ground truth depth during training, which is subject to availability.

Whereas **self-supervised** methods do not require this ground truth data. Typically, methods in this category predict a disparity map. Estimating the disparity map for each of the images in the stereo pair enables synthesizing the corresponding image via warping the original. For example, given the left input image, the right disparity map can be used to warp this input, to generate a hypothesis of the right view. Hence self-supervised learning methods focus on the reconstruction of images based on optimizing for photometric re-projection error by comparing the original input images and the synthesized predicted images. Ensuring these synthesized images appear similar to the original input indirectly optimizes the model to hone its disparity map output. This disparity can then be utilized to infer depth since depth is inversely proportional to disparity.

One of the earliest methods to formulate a self-supervised monocular depth estimation approach was [5]. They utilized a photometric consistency-based loss to train their deep learning model, in particular an L2 loss. In [5] the reconstruction loss is not fully differentiable, hence a Taylor approximation is performed to linearize the loss with respect to the model, which adds further complexity to the method. Moreover, Godard, Mac Aodha,

and Brostow [6] utilized a weighted sum of Multi-Structural Similarity Index (MS-SSIM) [7] and L1 distance as their photometric error. They also introduced two additional components as part of their loss term: the disparity smoothness loss computed between the left and right predicted disparity maps and the left-right disparity consistency loss. Both of which act as regularizers and ensure the output disparity is consistent with the corresponding one. These loss components from Godard, Mac Aodha, and Brostow [6] have become somewhat standard in self-unsupervised depth estimation.

Stereo image depth: This is where both the stereo pairs (left and right) are fed as an input to the model for depth estimation. Traditional stereo methods aim to generate a disparity map that minimizes an energy function, which essentially comprises of two terms: matching cost and smoothness regularization [8, 9]. Traditional methods tend to follow a four-step process for disparity estimation:

1. Feature extraction, from the two input images
2. Feature matching, between the two images
3. Computation of disparity
4. Refinement and post-processing of disparity

The first two steps construct a cost volume, which is simply a matrix that has aggregated the relevant features from the left and right image pipelines. Earlier deep learning-based methods in stereo matching simply replaced (1) & (2), which comprised of hand-designed feature extraction and matching with convolutional neural networks for doing high-dimensional feature matching [10]. Recent work advances on this methodology and focuses on improving stereo network architecture components. Chang and Chen [11] introduced a new network architecture, entitled “PSMnet,” which comprised of spatial pyramid pooling in their stereo matching network that consisted of 3D convolutions. This type of pooling module enables the network to encode hierarchical context information, which is difficult to attain from pixel intensities alone. Guo et al. [12] further developed on PSMnet, by innovating the way cost volume is aggregated in the network. Typically, this volume is aggregated by

concatenating left and right feature maps or via cross-correlation; Guo et al. [12] however, experimented with group-wise correlation. Unlike full correlation it is more efficient and still retains all the representations for conducting feature matching. The aforementioned methods are supervised learning based, and just like in monocular depth estimation self-supervised learning approaches are also utilized in stereo matching. These methods typically extend [6] to compute losses with respect to both input images or advance on the photometric loss component. Pilzer et al. [13] took a self-supervised approach by defining the photometric loss as an adversarial learning problem, where an additional network called the discriminator is utilized as a loss function [14].

Some of these aforementioned techniques in depth perception have also been utilized on surgical scenes to attain great performance for clinical applications. A recent challenge in minimal invasive surgery (MIS) entitled Stereo Correspondence and Reconstruction of Endoscopic Data, i.e., SCARED [15], had multiple submissions that utilized the work of Chan and Chen [11], Garg [5], and Goodfellow et al. [14] adapted to stereo inputs. These models achieved comparable performance on MIS data, to their natural scene counterparts. Another work by [16] utilized a self-supervised approach for depth perception on laparoscopic images. In particular they utilized an autoencoder model that took a pair of calibrated stereo images as inputs to predict their corresponding disparity maps. Similar to that of [5], however with the exception of stereo inputs as opposed to the monocular as done in the original. Furthermore, [17] trained a conditional generative adversarial network (pix2pix) [18], to generate a depth map from input monocular endoscopic images for colonoscopy. They originally train on synthetic images, due to lack of data availability and show robustness on transferring to real data during testing. They train their model in a joint adversarial and supervised manner. Furthermore, deep learning requires large quantities of training samples to train models, especially for self-supervised-based approaches for depth estimation. To that extent [19] created a software entitled “*VisionBlender*” that can be utilized for

generating synthetic data for minimally invasive surgical scenes including depth ground truths for training deep learning in surgery.

Navigation

Structure from Motion: Structure from Motion (SfM) is the process of extracting (3D) depth information and camera motion from a sequence of standard images (2D), typically a monocular video. Models designed for SfM need to estimate depth and camera pose between adjacent frames conjointly to understand the spatio-geometric relationship in the video. Traditional SfM methods rely on Bundle-Adjustment [20], where depth and camera motion for every view is conjointly optimized via Levenberg-Marquardt (LM) [21] algorithm, which is a form of iterative nonlinear least squares optimization. However this approach is successful only in limited scenarios since they are restricted to the regions overlapping in multiple views and also they fail to reconstruct textureless and reflective surfaces due to missing correspondences. Hence, deep learning tries to alleviate some of these shortcomings of conventional SfM. Zhou et al. [22] introduced “SfMLearner,” one of the first methods to train a depth prediction model from monocular video. They jointly trained two models, one that predicts depth and another that predicts camera pose between adjacent frames. Their model predicts depth and camera pose for the transition of current frame to the next, and current frame to the previous frame. For robustness to nonrigid scene motion, an additional mask was predicted that ignored the occluded pixels caused by motion in the scene. Similar to the self-supervised losses above, photometric per-pixel error is calculated between the predicted synthesized frame via warping and the unoccluded pixels of the ground truth to optimize the models. SfMLearner at the time was the state-of-the-art depth estimation from training solely on monocular videos. However, the sharpness of its predicted depth map was still an issue since missing finer structural details like edges of the object were not effectively penalized in the self-supervised loss.

The work in [23] further improved on this methodology by estimating multiple motion masks such as estimating object and camera motion separately and utilizing it to improve the depth prediction. Yin and Shi [24] further decomposed motion into rigid and nonrigid motion, through depth and optical flow. This enabled the network to be more robust to dynamic moving objects since the model can differentiate between the two types of motion in the scene. Rigid motion here is estimated by the camera motion estimation model and the non-rigid motion is estimated through processing optical flow maps and performing consistency checks between intermediate frames. More recently, Godard et al. [25] built upon the work of Zhou et al. [22], and showed that instead of averaging the reprojection loss between the past and future frames, taking the minimum of these reprojection losses attains superior performance. Furthermore, during training they also ignore stationary pixels with respect to ego-motion. Lastly they adopt the original [6] multi-scale structural similarity index measure (MS-SSIM) as part of their photometric loss.

Some examples of SfM in surgical scenes include [26] where the authors combined information from monocular videos captured in both structured light and white light with traditional SfM for surface reconstruction. In particular they utilize SURF features [27] detection conjointly with Lucas-Kanade optical flow estimation [28] to reconstruct shape. They assumed the surface of the organ system was rigid in a small time frame, therefore, making it possible to estimate the relative camera position via singular value decomposition (SVD) (provided enough correspondences were measured between the two consequent frames). Another example [29] designed a deformable structure from motion, where they fuse inertial motion captured by MEMS sensors together with vision, for more robust camera pose estimation. They fuse these two information modalities via a novel upgraded version of Unscented Kalman Filter. They follow the standard SfM pipeline for 3D reconstruction with the exception of accounting for tissue deformation via Gaussian mixture models for clustering pixels with coherent motion between video frames.

There is still limitation in the uptake of deep learning for structure from motion in surgical scenes, especially since deformable reconstruction is still an open research question. Furthermore estimating camera pose is challenging when the scene has dynamic moving objects and deformable scenes. Hence, improvement in methods robust to these challenges can result in more modern deep learning methods utilized for solving this research question. In particular, an interesting study by Tukra et al. [30] trained a see-through vision convolutional neural network, for removing occlusions in monocular endoscopic video. Potentially this can be further utilized to reduce the impact of highly dynamic objects within the scene to enable more robust camera registration and localization pipeline and indirectly the SfM 3D reconstruction. An example of this type of “see-through vision” is shown in Fig. 2.

Simultaneous Localization and Mapping (SLAM): This takes the information from step 1 and 2 in Fig. 1 to enable the robot to localize itself in the environment. Hence, this section covers step 3 of Fig. 1. SLAM shares some of the characteristics of SfM such as jointly estimating camera pose and depth map of the scene via features detected across multiple frames and both utilize Bundle Adjustment [20]. SLAM however, takes this process one step further to build a map of the environment to enable robots to navigate and track certain key points of interest in it. Hence, the goal of SLAM can be distilled to estimating positions of the robot onto a map, thereby utilizing multiple sensory data, depth information, camera pose, etc., to achieve it. SLAM bridges the gap between SfM for scene understanding to proprioception for robotics.

Traditional structure of SLAM is defined by the following steps:

- Feature detection: of points of interest within the scene
- Feature matching: matching the detected features between adjacent images and tracking them
- Optimization: updating these features to be robust to adversarial effects like motion, blur, etc.

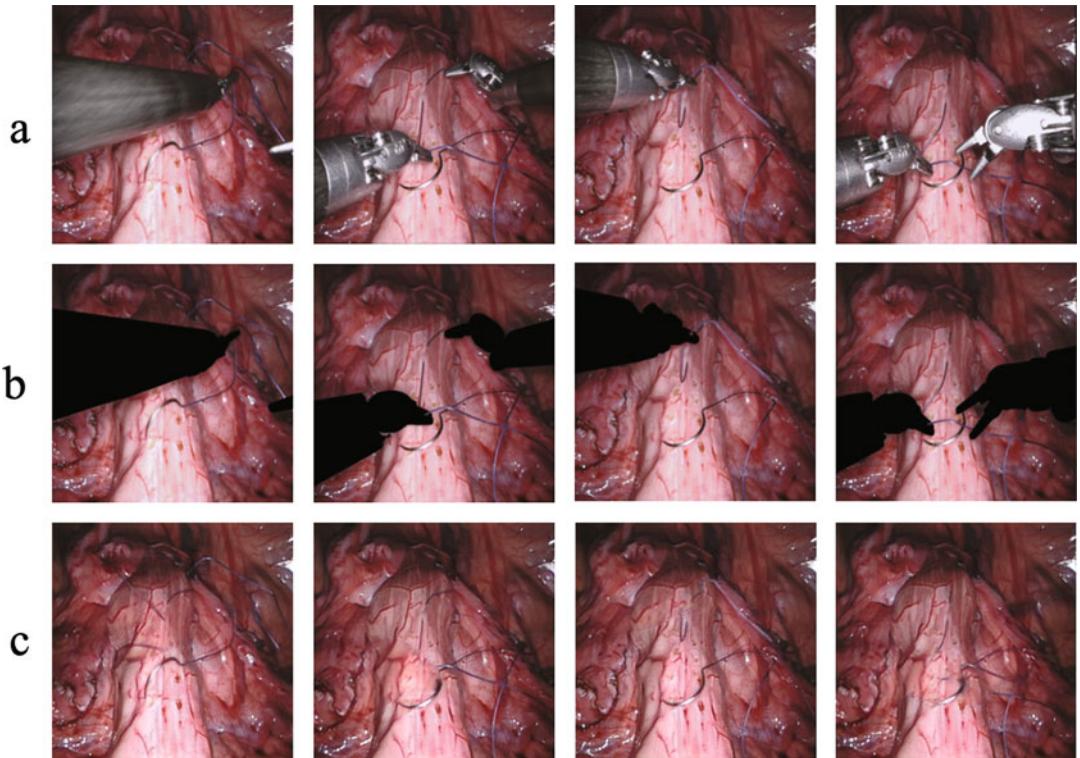


Fig. 2 Showing the output of [30] given a monocular video from a surgical scene. Where (a) are the individual video frames from the original source video with the robotic tool as the occlusions, (b) the robotic tool occlusion

mask, and (c) the reconstructed video frames by Tukra et al. [30] where the tool occlusions have been removed. (Figure reproduced from [30])

- Loop closure detection: an algorithm that processes previously visited areas and/or information to define a map

Combining all of the points above creates a full working SLAM system, and typically research focuses on improving one or multiple of these individual tasks using methods from modern machine learning. Additionally, SLAM has been combined with sensory inputs to attain better results, and such sensors fall into two categories: proprioceptive and exteroceptive. Proprioceptive sensors enable the robot to map its own trajectory like gyroscopes, inertial measurement units (IMUs), etc. Exteroceptive sensors enable the robots to understand the external world around them, such that environmental information is registered. Examples of such sensors include cameras (mono, stereo), LIDARs, etc. In this particular section we will discuss SLAM purely focusing on computer vision, camera-based

inputs as this is where machine learning plays a huge role in SLAM (context understanding).

One of the earliest works done to prove SLAM is indeed possible via a single camera as a data source was MonoSLAM [31]. Here the authors took a filtering approach, which involves estimating a probability density function over the robot's current position with respect to the features detected on images, via Extended Kalman Filter (EKF). Peter Mountney et al. [32] extended this EKF SLAM method to stereoscopic videos and estimated soft tissue deformation as periodic motion of the respiratory system. Grasa et al. [33] experimented with Andrew J. Davison et al. [31]'s method to see the efficacy of EKF SLAM on hernia repair surgeries. The most popular form of SLAM that is also available on open source programs like OpenCV today is ORB-SLAM [34], that utilizes ORB features to perform keypoint matching between intermediate frames. Song et al. [35] introduced MIS-SLAM,

which further built on ORB-SLAM by using the full potential of compute power, i.e., both CPU and GPU for computation, and thereby achieving high performance even from ORB features. Dense SLAM was conducted on the GPU and ORB-SLAM was conducted on the CPU and later the two estimates were fused for greater performance.

Surgical Tool Tracking

Tool tracking in the context of robotic surgery entails the estimation of the translation and rotation of an instrument as it moves with six degrees of freedom (6DoF) in the surgical environment. In the literature, frameworks that fuse kinematic and visual information have been proposed [36, 37]. However, kinematic information is not always available and in addition these methods rely on the transformation from the laparoscope to one of the actuators to be precisely estimated and continuously updated [38]. Therefore, vision-based tool tracking frameworks have been developed that are highly attractive since they rely on laparoscopic video data, which is directly available and do not require changing the operating theater or the design of surgical tools. This category of methods includes feature-based frameworks, and end-to-end tool pose estimation using deep learning [39, 40]. The main issue with current deep learning frameworks is the lack of generalizability. More specifically, if the camera parameters change, the model needs to be retrained or fine-tuned. Recent approaches aim at overcoming this limitation [41].

Feature-based tool tracking frameworks can be divided into marker-less and marker-based methods. Marker-less frameworks focus on tracking surgical tools using features that are naturally present on the surgical tool [37, 42, 43]. These features depend on the specific surgical tool; hence a new set of features must be extracted every time a new tool is used. Conversely, marker-based frameworks are less restrictive since they can be used by simply attaching a marker to a surgical tool, making

them an attractive solution for tracking. In robot-assisted Minimally Invasive Surgery (MIS), depending on the shape of the tool, planar and cylindrical markers have been used. Planar markers have been used to track imaging probes with planar surfaces [44, 45]. Since in robotic surgery most of the surgical tools are cylindrical objects, frameworks for tracking cylindrical markers have been developed [46–51]. Although tracking of planar markers can achieve high accuracy, the tracking error of cylindrical marker frameworks is still in the millimeter range, which is too large for applications that require sub-millimeter accuracy such as neurosurgery.

Haptic Feedback and Tissue Interaction Sensing

Surgeons heavily rely on haptic feedback when they are operating as it can enable them to identify buried tissue structures or stiffer regions during palpation, as well as safely control tool–tissue interactions, preventing iatrogenic injuries. To achieve haptic feedback in robotic surgery, extensive research has been carried out on integrating force sensors on the tip of surgical instruments. However, the additional complexity due to biocompatibility, sterilization, and space constraints and the cost associated with external sensing, limit the clinical applicability of these methods [52]. Technological advancements in sensing and actuation have led to the development of approaches based on mechatronics [53] as well as pneumatic [54] and hydraulic [55] systems. Recently, neural networks have been used to estimate the inverse dynamics of the da Vinci surgical robot, which enables estimation of the external torques/forces acting on the joints of the robot [56]. In another attempt to gain force sensing with the da Vinci robot, a deep learning framework has been proposed for end-to-end force estimation [57]. The model is trained using data collected by both moving an instrument in free space and by palpating a tissue phantom that has an embedded force sensor for ground truth (Fig. 3).

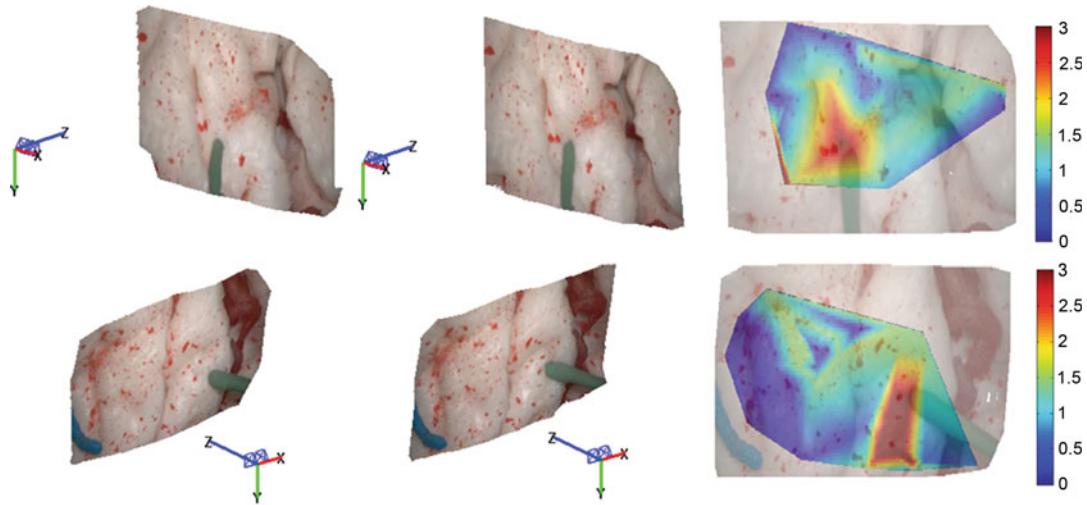


Fig. 3 Vision-based tissue deformation recovery. Estimated 3D structure before tool–tissue interaction (first column); estimated 3D structure during tool–tissue

interaction which causes tissue deformation (second column); deformation heatmaps overlapped on the images (third column) [37]

As an alternative, vision can provide an indication of exerted forces. Vision-based approaches for force estimation, which are based on the use of the existing endoscopic cameras, are particularly appealing as they do not require the introduction of additional equipment to what is often an already crowded operating room, and allow seamless integration into the existing surgical workflow. For this purpose, stereo depth estimation has been combined with surface mesh analysis based on spring-damper models [58] to recover force information. The vision-based framework proposed in [59] used quasi-dense 3D stereo reconstruction and soft tissue tracking to recover tissue deformation. The surface displacement has been combined with biomechanical modeling to estimate applied forces during tool–tissue interaction. A recurrent neural network architecture has been proposed in [60] to map soft-tissue deformation and tool data to interaction forces. This work has been enhanced in [61] with the incorporation of a Long-Short Term Memory (LSTM) network. Recently, a force estimation model based on CNN and LSTM networks has been proposed using both, the spatiotemporal information present in video sequences and the temporal structure of tool data such as tool-tip trajectory and grasping status [62]. In the above literature, the vision-

based estimation of forces has mainly been studied on simple tasks, such as pushing, whereas surgical tasks that result in complex interactions, such as grasping, have not been explored yet.

Advanced Visualization with Augmented Reality

Augmented Reality (AR) enables the surgeon to visualize subsurface anatomical structures such as tumors, by overlaying virtual objects on the endoscopic scene. For realistic results, AR visualization needs to be capable of seamlessly integrating multimodal images of dynamic anatomical structures while attaining a high level of spatiotemporal accuracy. This alignment is achieved using registration methods that either rely on external devices, such as optical or electromagnetic tracking systems [63] or focus purely on the processing of intraoperative images [64]. Although the latter category of registration methods has greater potential for clinical deployments as it does not require additional hardware, it is not yet widely used in surgery. This is due to the highly complex deformations of the anatomical structures during the operation due to tissue motion as well as topological changes during resections. For

successful registration of multimodal images, similarity measures have been proposed, which are invariant to local changes in brightness and contrast [65, 66]. Recently, unsupervised deep learning-based methods have been proposed for image registration based on the optimization of a loss function, which represents image similarity [67]. However, these methods cannot deal with multiple modalities or significant data variability. To overcome this limitation, deep learning frameworks have been proposed which, instead of modeling intensity correlations between images, optimize alternative correspondence measures such as the overlap of corresponding image areas, or spatial transformation models [68]. To guarantee successful image alignment, direct approaches to initialization have been proposed [69, 70]. Another category of image registration methods is based on the detection and identification of anatomical landmarks. In this case, deep learning models have been designed to produce a pose invariant latent representation of the appearance of landmarks, which can be used to establish image correspondences [71–73]. However, the performance of these methods has not been evaluated yet under challenging conditions with significant image pose differences.

In the last decade, there has been a great effort to bring mixed reality (MR) into the operating room to assist surgeons intraoperatively. MR “mixes” virtual and real objects, thus allowing, for example, a surgeon to see a virtual tumor inside a real patient’s body. In addition, it allows a surgeon to consult data when and where needed, making it a valuable tool for intraoperative decision-making. Using the Microsoft HoloLens, previous research has mainly focused on projecting a virtual 3D model into a patient’s body [74–76] or just above it to avoid obstructing a surgeon’s line of sight [77]. Besides 3D virtual objects, a few studies [78] have used MR to display surgical data through virtual 2D screens, which can also display crucial data to the surgeon from multiple imaging modalities. However, in the above studies, the interactions between a surgeon and the virtual world have been restricted to moving, scaling, or rotating the virtual objects. Recently, a MR visualization platform has been developed for the HoloLens, which

advances existing platforms by integrating multimodal data for intraoperative surgical guidance and incorporating novel interactive functionalities [79]. The visualization components include a 3D organ model, volumetric data, and tissue morphology captured with intraoperative imaging modalities. The introduced functionalities allow the surgeon to customize and interact with virtual objects, namely, scrolling through volumetric data and transparency adjustment of the objects. A pilot study verified the usability of this platform in the operating theater. Despite the above progress, we are still on an early stage to achieve the goal of bringing MR into a standard operating room to assist surgeons intraoperatively. In addition, little attention has been paid to get feedback from surgeons participating in MR experiments [80].

Robot-Assisted Task Execution

Current research on autonomous robotic task execution has mainly focused on robot-assisted ultrasound elastography [81], motion compensation in cardiovascular surgery [82], autonomous tissue dissection [83], brain ablation [84], and cochlear implant installation [85]. Robot-assisted tissue scanning with imaging probes under image guidance has attracted significant interest. Current approaches to robot-assisted tissue scanning with imaging probes have focused on applications using Ultrasound and probe-based Confocal Laser Endomicroscopy (pCLE) as imaging modalities [86, 87]. This is because the above modalities should firmly touch the tissue surface while closely following the tissue motion to capture good quality imaging data (Fig. 4).

Current research on autonomous scanning with pCLE has focused on optimizing the contact between the imaging probe and the tissue. For this purpose, force feedback from force sensors and stereo vision for pose estimation have been used in [86]. The visual servoing framework proposed in [89] employed Optical Coherence Tomography (OCT) to estimate the distance between the probe and the tissue to control the probe during scanning. Reinforcement learning and image blur characteristics have been used to

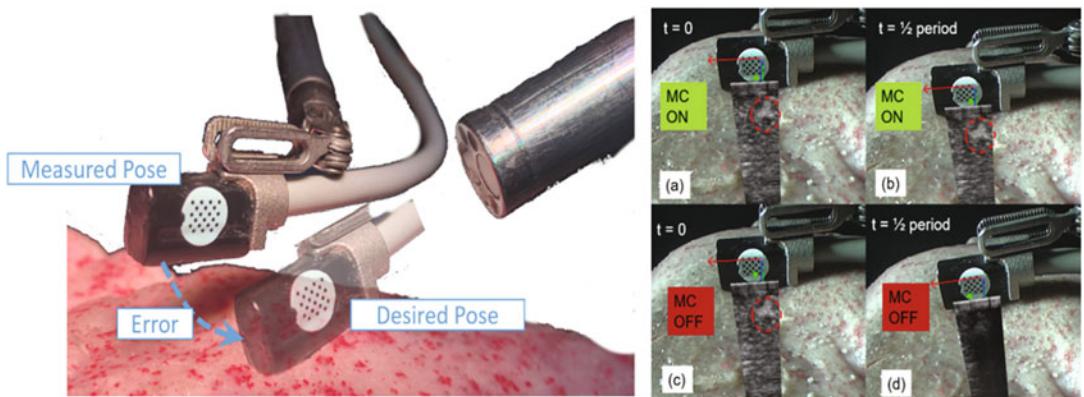


Fig. 4 Autonomous Tissue Scanning under Free-Form Motion. (Left) Graphical representation of visual servoing. (Right) Robot-assisted tissue scanning with and without

classify the position of the probe as too close, too far, or at the right distance from the tissue [90]. Endomicroscopic information has also been used in [91] to control the distance between the probe and the tissue applying the Crete-Roffet Blur Metric (CRBM) along with fuzzy logic. To expand the small field of view captured by pCLE probes and enhance visualization, image mosaics have been generated [92, 93].

Previous studies on autonomous Ultrasounds scanning have focused on image-guided needle insertion [94], Focused Assessment with Sonography for Trauma (FAST) [95], tumor detection [87], and vessel tracking [96]. In the above, visual servoing frameworks information either from the Ultrasounds data [97] or the endoscopic camera [98] has been used as visual feedback for probe positioning or sweeping scanning. Recently, force feedback has also been included for Ultrasounds-guided flexible needle insertion with haptic feedback [99]. For applications where the robot trajectory needs to be planned in advance, the 3D structure has been recovered using stereo cameras [87], RGB-D cameras [100], or combination of RGB-D and preoperative MRI images [101].

The above approaches to robot-assisted local tissue scanning rely on the assumption that the tissue is static or moving with periodic motion. The visual servoing framework for autonomous tissue scanning proposed in [88] advances state-of-the-art autonomous tissue scanning methods

by eliminating the requirement for learning the tissue motion before scanning and dealing with free-form tissue motion in real-time. It optimizes probe-tissue contact by tracking the motion of the tissue surface and updating the desired scanning trajectory to follow the tissue motion and control the movement of the robotic arm.

Context-Aware Decision Support

Recent advances in biophotonics have allowed intraoperative tissue characterization with the advantages of being noninvasive or minimally invasive. For instance, techniques such as probe-based Confocal Laser Endomicroscopy (pCLE) allow for real-time morphological imaging at sub-cellular resolution and have demonstrated efficacy in distinguishing neoplastic lesions from normal tissue [102]. Furthermore, intelligent surgical devices such as the “iKnife” and Laser Desorption Ionization Mass Spectrometry (LDI-MS) enable real time tumor phenotyping in the operating room for tissue characterization that corresponds with histopathology [103]. Raman spectroscopy (RS) [104] can analyze the tissue *in vivo*, *in situ* at a microscopic scale, and provide information about its histochemical state. Recently, the use of hyperspectral imaging has shown promise in neurosurgical oncology for intraoperative margin delineation at macroscopic scale [105].

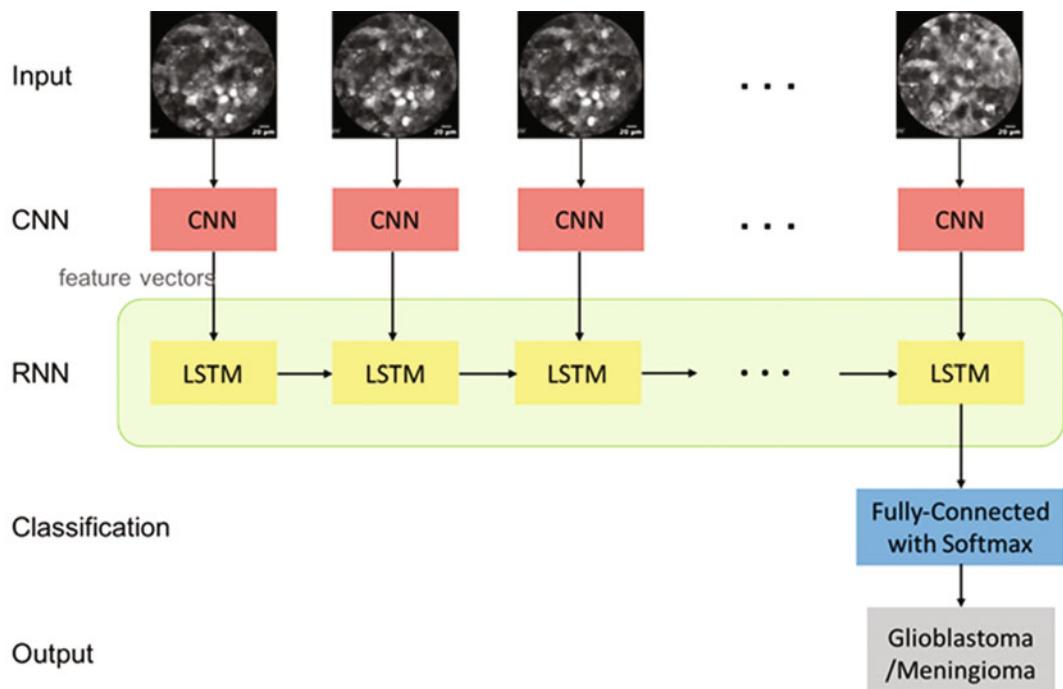


Fig. 5 Context aware decision support in neurosurgical oncology based on video classification of Glioblastoma and Meningioma endomicroscopic data [113]

However, the interpretation of the information captured with the above techniques remains challenging, particularly for surgeons who do not themselves routinely review histopathology or biochemical data. There is also significant intra- and interclass variability in the captured tissue characteristics. Furthermore, even among experts, the diagnosis can be examiner-dependent, leading to considerable interobserver variability. Hence, automatic tissue characterization, based on a database of previously annotated data by expert physicians with diagnosis confirmed by histology, would support the surgeon in establishing diagnosis and guide autonomous robotic tissue scanning to focus locally on pathological areas (Fig. 5).

Computer-aided diagnosis (CAD) systems designed for the analysis of pCLE data are composed of three main steps, namely, visual feature extraction, feature encoding, and supervised classification. A content-based image and video retrieval framework based on the Bag-of-visual Words (BoW) approach has been proposed in [106] for the differentiation of neoplastic and

benign colorectal polyps. Clinical evaluation of this framework presented in [107] shows that automated classification of pCLE videos of colonic polyps achieves high performance, comparable to off-line diagnosis of pCLE videos established by expert endoscopists. The above video retrieval framework has been extended in [108] to include high-level knowledge for the pathological interpretation of pCLE images. For that purpose, binary semantic concepts commonly used by expert endoscopists to diagnose pCLE videos of colonic polyps have been extracted and used as additional information that complements the visual outputs of the retrieval framework.

Wan et al. [109] used an efficient feature encoding scheme based on codeword proximity and a support vector machine (SVM) classifier to classify from pCLE images two types of commonly diagnosed brain tumors, namely, Glioblastomas and Meningiomas. This work has been extended in [110] by exploring more encoding schemes for data description and using a majority

voting-based classification scheme for video classification. To further improve the retrieval accuracy, approaches to content-based image retrieval of pCLE data have focused on learning discriminative visual features. In [111], a MultiView Multi-Modal Embedding (MVMME) framework has been proposed to learn in a supervised way discriminative features of pCLE videos by exploiting both mosaics and histology images. This work has been extended in [112] where an unsupervised multimodal graph mining (UMGM) approach has been proposed to learn the discriminative features for pCLE mosaics of breast tissue.

Recently, an efficient representation of the context of pCLE data has been proposed in [113] by exploring CNN models with different tuning configurations to classify brain tissue into Glioblastoma and Meningioma. Furthermore, a video classification framework based on the combination of convolutional layers with long-range temporal recursion has been developed to estimate the probability of each tumor class. Deep learning models have also been designed for super-resolution in endomicroscopy [114, 115].

Hyperspectral Imaging (HSI) captures spectral and spatial data beyond the limited electromagnetic bands perceived by the human eye. It has been proven that the interaction between electronic radiation and tissue carries useful information for tumor detection in surgery. Although medical HSI data bears rich information, it is characterized by high-dimensionality and difficulty to interpret for clinicians as it generates a temporal flow of 3D information that cannot be simply displayed in an intuitive fashion. In addition, the inter-patient spectral variability and the limited number of samples, makes the processing of HSI data challenging.

The processing of HSI entails mainly the pixel-wise classification of the data based on its spectral information. For this purpose, Machine Learning techniques based on Linear Discriminant Analysis (LDA), decision trees, random forest (RF), and kernel-based methods have been developed. For gastric cancer detection, quadratic SVMs have been used to classify fat, muscle, and tumor areas on HSI data with high accuracy [116]. In [117], the recursive divergence method has been used to

select the most significant wavelengths in the spectral range, which have been used to detect colon cancer on *in vivo* data. To identify cancerous and noncancerous areas on breast tissue [118], the Fourier coefficient selection method has been used to extract features from HSI data, followed by dimensionality reduction using the Minimum Redundancy Maximum Relevance method. To classify the two tissue states, the SVM classifier was used with the radial basis function (RBF) kernel.

The ensemble linear discriminant analysis (LDA) has been successfully used in [119] to delineate the boundaries between the normal and cancerous tissue on head and neck tissue samples. LDA has been combined with PCA for dimensionality reduction to identify malignant changes in the oral cavity [120] and in particular to discriminate between healthy, hyperplastic, dysplastic, and squamous cell carcinoma (SCC) tissue. For the processing of laryngeal HSI data [121], rigid image-to-image registration based on normalized cross correlation (NCC) was used as a pre-processing step to deal with misalignment due to cardiac motion as well as noise removal based on the minimum noise fraction (MNF) transformation. To identify healthy and cancerous tissue areas, the random forest (RF) classifier was employed.

To facilitate the semantic segmentation of brain HSI data, the Fixed Reference T-distributed Stochastic Neighbors (FR-t-SNE) method has been proposed in [122] to reduce the dimensionality of the data volume and generate high contrast images for accurate brain tumor detection. In [123], supervised classification based on a combination of a SVM, PCA, and KNN algorithms was fused with the K-means classifier for unsupervised segmentation, through a majority voting procedure, to delineate the margin of brain tumors.

Recently, deep learning has been used to extract high-level spatial features from HS data and classify different tissue areas. CNN-based architectures have been developed to discriminate between cancer and normal tissue in Head and Neck surgery [124, 125], outperforming traditional ML techniques. Promising results have also been reported for the performance of deep learning approaches on the classification of

glioblastoma tumors [126, 127]. To take into account the inter-patient variability, a leave-one-patient-out cross-validation approach was followed. Despite the good performance, extensive validation of deep learning techniques is still required on large medical datasets. In addition, clinical validations of the developed systems should be carried out to assess their performance.

Outlook

Robotics has huge potential in being widely adopted in surgery, due to improved sensing, machine learning pipelines for perception, and deep learning for enhanced feature understanding. Endowing cognition in such robotic tools such that they are robust to the ever-changing scenarios presented in a surgical scene. Especially since the use of AI allows learning a shared representation from multiple sources of data captured through sensors such that robots can have an improved representation of their surrounding environment. Enabling them to comprehend the complexity of the environment and model it for decision-making, greater precision, and eventual automation of surgical tasks. Therefore, paving the foundation to change the surgical practice as a whole and moving it into the new age of intelligence, comprising advancements in (and not limited to) technologies for imaging, navigation, and robotic intervention.

However due to the general complexity of the dynamic surgical environment, such AI methods still require improvement in generalizability, to accommodate transferring its learnt knowledge to other unobserved tasks while reducing its hunger for new data. One consideration for the future is understanding the ethical and moral deployment of such autonomous/cognitive robots in surgery. It is vital that research into human behavior, ethical and legal implications in the use of autonomous robotics, is further advanced in parallel to technological advances in AI. As robots definitely bring divergent advantages to the field, one cannot neglect how they may be perceived in the eyes of a patient. Safety being the top priority, the rules and regulations behind privacy, accountability in

presence of errors, ethical decision-making in whether robotic surgery should even be conducted or an alternative should be performed, has to be updated alongside robotics. We still have a distance to go in endowing robotics with the sensory and cognitive ability of surgeons. However, the current ongoing research in fusion of AI and robotics is certainly moving toward that direction.

References

- Yang GZ, Cambias J, Cleary K, Daimler E, Drake J, et al. Medical robotics – regulatory, ethical, and legal considerations for increasing levels of autonomy. *Sci Robot*. 2017;2:eaam8638.
- Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, editors. Advances in neural information processing systems 27. Curran Associates; 2014. p. 2366–74. <http://papers.nips.cc/paper/5539-depth-map-prediction-from-a-single-image-using-a-multi-scale-deep-network.pdf>.
- Cao Y ZW, Shen C. Estimating depth from monocular images as classification using deep fully convolutional residual networks; 2018. p. 3174–82. <https://ieeexplore.ieee.org/document/8010878/authorsauthors>
- Fu H, Gong M, Wang C, Batmanghelich K, Tao D. Deep ordinal regression network for monocular depth estimation. In: Proceedings of CVPR. 2018. p. 2002–11.
- Garg R, G VKB, Reid ID. Unsupervised CNN for single view depth estimation: geometry to the rescue. European Conference on Computer Vision (ECCV). 2016;abs/1603.04992. <http://arxiv.org/abs/1603.04992>
- Godard C, Mac Aodha O, Brostow GJ. Unsupervised monocular depth estimation with left-right consistency. In: IEEE conference on computer vision and pattern recognition (CVPR). 2017. <http://visual.cs.ucl.ac.uk/pubs/monoDepth/>
- Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process*. 2004;13(4):600–12. <https://doi.org/10.1109/TIP.2003.819861>.
- Woodford OJ, Torr PHS, Reid ID, Fitzgibbon AW. Global stereo reconstruction under second-order smoothness priors. *IEEE Trans Pattern Anal Mach Intell*. 2009;31(12):2115–28. <https://doi.org/10.1109/TPAMI.2009.131>.
- Hirschmuller H. Stereo processing by semiglobal matching and mutual information. *IEEE Trans Pattern Anal Mach Intell*. 2008;30(2):328–41.

10. Žbontar J, LeCun Y. Stereo matching by training a convolutional neural network to compare image patches. *J Mach Learn Res.* 2016;17(1):2287–318.
11. Chang JR, Chen YS. Pyramid stereo matching network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE; 2018. p. 5410–8.
12. Guo X, Yang K, Yang W, Wang X, Li H. Group-wise correlation stereo network. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). IEEE; 2019. p. 3273–82.
13. Pilzer A, Xu D, Puscas M, Ricci E, Sebe N. Unsupervised adversarial depth estimation using cycled generative networks. In: 2018 international conference on 3D vision (3DV). IEEE; 2018. p. 587–95.
14. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger KQ, editors. Advances in neural information processing systems, vol. 27. Curran Associates; 2014. <https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf>.
15. Allan M, Jonathan McLeod A, et al. Stereo correspondence and reconstruction of endoscopic data challenge. CoRR. 2021. abs/2101.01133.
16. Xu K, Chen Z, Jia F. Unsupervised binocular depth prediction network for laparoscopic surgery. *Comput Assist Surg.* 2019;24(Suppl 1):30–5.
17. Rau A, Edwards PJE, Ahmad OF, et al. Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. *Int J CARS* 14, 1167–1176 (2019). <https://doi.org/10.1007/s11548-019-01962-w>.
18. Isola P, Zhu JY, Zhou T, Efros A. Image-to-image translation with conditional adversarial networks. CVPR; 2017.
19. Cartucho J, Tukra S, Li Y, Elson DS, Giannarou S. VisionBlender: a tool to efficiently generate computer vision datasets for robotic surgery. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*; 2020. p. 1–8.
20. Triggs B, McLauchlan PF, Hartley RI, Fitzgibbon AW. Bundle adjustment – a modern synthesis. In: Proceedings of the international workshop on vision algorithms: theory and practice. ICCV ‘99. Berlin/Heidelberg: Springer; 1999. p. 298–372.
21. Moré J. The Levenberg-Marquardt algorithm: implementation and theory. In: Watson GA, editor. Numerical analysis. Vol. 630 of Lecture notes in mathematics. Berlin/Heidelberg: Springer; 1978. p. 105–16. <https://doi.org/10.1007/BFb0067700>.
22. Zhou T, Brown M, Snavely N, Lowe DG. Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). IEEE; 2017.
23. Fragkiadaki A, Seybold B, Schmid C, Sukthankar R, Vijayanarasimhan S, Ricco S. Self-supervised learning of structure and motion from video. arxiv. 2017;2017. <https://arxiv.org/abs/1704.07804>
24. Yin Z, Shi J. GeoNet: unsupervised learning of dense depth, optical flow and camera pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). IEEE; 2018.
25. Godard C, Mac Aodha O, Firman M, Brostow GJ. Digging into self-supervised monocular depth prediction. Proceedings of ICCV 2019, October.
26. Lin J, Clancy NT, Hu Y, Qi J, Tatla T, Stoyanov D, Maier-Hein L, Elson DS. Endoscopic depth measurement and super-spectral-resolution imaging. In: Medical image computing and computer assisted intervention – MICCAI 2017 – 20th international conference, Quebec City, QC, Canada, September 11–13, 2017, Proceedings, Part II. Springer; 2017. p. 39–47.
27. Bay L. SURF: speeded up robust features. In: Computer vision – ECCV 2006. Berlin/Heidelberg: Springer; 2006. p. 404–17.
28. Lucas B, Kanade T. An iterative image registration technique with an application to stereo vision. In: Proceedings of the international joint conference on artificial intelligence. Kaufmann; 1981. p. 674–9.
29. Giannarou S, Zhang Z, Yang G. Deformable structure from motion by fusing visual and inertial measurement data. 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2012, pp. 4816–4821. <https://doi.org/10.1109/IROS.2012.6385671>.
30. Tukra S, Marcus HJ, Giannarou S. See-through vision with unsupervised scene occlusion reconstruction. *IEEE Trans Pattern Anal Mach Intell.* 2021. <https://doi.org/10.1109/TPAMI.2021.3058410>. Epub ahead of print.
31. Davison AJ, Reid ID, Molton ND, Stasse O. MonoSLAM: real-time single camera SLAM. *IEEE Trans Pattern Anal Mach Intell.* 2007;29: 1052–2007.
32. Mountney P, Stoyanov D, Davison AJ, Yang G-Z. Simultaneous stereoscope localization and soft-tissue mapping for minimal invasive surgery. In: Medical image computing and computer-assisted intervention – MICCAI 2006, 9th international conference, Copenhagen, Denmark, October 1–6, 2006, Proceedings, Part I. Springer; 2006. p. 347–54.
33. Grasa ÓG, Bernal E, Casado S, Gil I, Montiel JMM. Visual SLAM for handheld monocular endoscope. *IEEE Trans Med Imaging.* 2014;33(1):135–46. <https://doi.org/10.1109/TMI.2013.2282997>.
34. Mur-Artal R, Montiel J, Tardós J. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Trans Robot.* 2015;31(5):1147–63.
35. Song J, Wang J, Zhao L, Huang S, Dissanayake G. MIS-SLAM: real-time large-scale dense deformable SLAM system in minimal invasive surgery based on heterogeneous computing. *IEEE Robot Automat Lett.* 2018;3(4):4068–75. <https://doi.org/10.1109/LRA.2018.2856519>.
36. Hao R, Ozguner O, Cavusoglu MC. Vision-based surgical tool pose estimation for the Da Vinci® robotic surgical system. In: 2018 IEEE/RSJ

- international conference on intelligent robots and systems (IROS). IEEE; 2018. p. 1298–305.
37. Ye M, Zhang L, Giannarou S, Yang GZ. Real-time 3d tracking of articulated tools for robotic surgery. In: International conference on medical image computing and computer-assisted intervention. Springer; 2016. p. 386–94.
38. Shao J, Luo H, Xiao D, Hu Q, Jia F. Progressive hand-eye calibration for laparoscopic surgery navigation. In: Computer assisted and robotic endoscopy and clinical image-based procedures. Springer; 2017. p. 42–9.
39. Kendall A, Grimes M, Cipolla R. Posenet: A convolutional network for realtime 6-dof camera relocalization. In: Proceedings of the IEEE international conference on computer vision. IEEE; 2015. p. 2938–46.
40. Mahendran S, Ali H, Vidal R. 3d pose regression using convolutional neural networks. In: Proceedings of the IEEE international conference on computer vision workshops. IEEE Computer Society; 2017. p. 2174–82.
41. Facil JM, Ummenhofer B, Zhou H, Montesano L, Brox T, Civera J. Camconvs: camera-aware multi-scale convolutions for single-view depth. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. IEEE Computer Society; 2019. p. 11826–35.
42. Reiter A, Allen PK, Zhao T. Articulated surgical tool detection using virtually rendered templates. In: Computer assisted radiology and surgery (CARS). 2012. p. 1–8.
43. Reiter A, Allen PK, Zhao T. Feature classification for tracking articulated surgical tools. In: MICCAI. Springer; 2012. p. 592–600.
44. Zhan J, Cartucho J, Giannarou S. Autonomous tissue scanning under free-form motion for intraoperative tissue characterisation. In: ICRA. IEEE; 2020. p. 11147–54.
45. Ma L, Wang J, Kiyomatsu H, Tsukihara H, Sakuma I, Kobayashi E. Surgical navigation system for laparoscopic lateral pelvic lymph node dissection in rectal cancer surgery using laparoscopic-vision-tracked ultrasonic imaging. *Surg Endosc*. 2020 Nov 13. <https://doi.org/10.1007/s00464-020-08153-8>. Epub ahead of print. PMID: 33185764.
46. Jayaratne UL, McLeod AJ, Peters TM, Chen ECS. Robust intraoperative US probe tracking using a monocular endoscopic camera. In: MICCAI. Springer; 2013. p. 363–70.
47. Jayaratne UL, Chen EC, Moore J, Peters TM. Robust, intrinsic tracking of a laparoscopic ultrasound probe for ultrasound augmented laparoscopy. *IEEE Trans Med Imaging*. 2018;38(2):460–9.
48. Zhang L, Ye M, Chan PL, Yang GZ. Real-time surgical tool tracking and pose estimation using a hybrid cylindrical marker. *IJCARS*. 2017;12(6):921–30.
49. Gadwe A, Ren H. Real-time 6dof pose estimation of endoscopic instruments using printable markers. *IEEE Sensors J*. 2018;19(6):2338–46.
50. Zhou D, Dong X, Zhang F, Chen W. A match method of encircled marker points on external store model. In: ICCSE. IEEE; 2019. p. 533–8.
51. Huang B, Tsai YY, Cartucho J, Vyas K, Tuch D, Giannarou S, Elson DS. Tracking and visualization of the sensing area for a tethered laparoscopic gamma probe. *IJCARS*. 2020;15(8):1389–97.
52. Marcus HJ, Payne CJ, Hughes-Hallett A, Gras G, Leibrandt K, Nandi D, Yang GZ. Making the leap: the translation of innovative surgical devices from the laboratory to the operating room. *Ann Surg*. 2015;263:1077.
53. Naghibi H, Hoitzing WB, Stramigioli S, Abayazid M. A flexible endoscopic sensing module for force haptic feedback integration. In: 2018 9th Cairo international biomedical engineering conference. Piscataway: IEEE; 2018. p. 158–61.
54. Hodgson S, Tavakoli M, Lelevé A, Tu Pham M. High-fidelity sliding mode control of a pneumatic haptic teleoperation system. *Adv Robot*. 2014;28: 659–71.
55. Ogawa K, Ohnishi K, Ibrahim Y. Development of flexible haptic forceps based on the electrohydraulic transmission system. *IEEE Trans Ind Inform*. 2018;14:5256–67.
56. Yilmaz2020. Neural network based inverse dynamics identification and external force estimation on the da Vinci research kit.
57. Tran2020. A deep learning approach to intrinsic force sensing on the da vinci surgical robot.
58. Kim W, Seung S, Choi H, Park S, Ko SY, Park JO. Image-based force estimation of deformable tissue using depth map for single-port surgical robot. In: 12th international conference on control, automation and systems (ICCAS). IEEE; 2012. p. 1716–9.
59. Giannarou S, Ye M, Gras G, Leibrandt K, Marcus HJ, Yang G-Z. Vision-based deformation recovery for intraoperative force estimation of tool-tissue interaction for neurosurgery. *Int J Comput Assist Radiol Surg*. 2016;11(6):929–36.
60. Aviles AI, Marban A, Sobrevilla P, Fernandez J, Casals A. A recurrent neural network approach for 3d vision-based force estimation. In: 4th international conference on image processing theory, tools and applications (IPTA). IEEE; 2014. p. 1–6.
61. Rivero AIA, Alsaled SM, Hahn JK, Casals A. Towards retrieving force feedback in robotic-assisted surgery: a supervised neuro-recurrent-vision approach. *IEEE Trans Haptics*. 2017;10(3):431–43.
62. Marban A, Srinivasan V, Samek W, Fernández J, Casals A. A recurrent convolutional neural network approach for sensorless force estimation in robotic surgery. *Biomed Signal Process Control*. 2019;50: 134–50.
63. Koivukangas T, Katisko JP, Koivukangas JP. Technical accuracy of optical and the electromagnetic tracking systems. *SpringerPlus*. 2013;2(1):90.
64. Liao R, Zhang L, Sun Y, Miao S, Chefd C. A review of recent advances in registration techniques applied

- to minimally invasive therapy. *IEEE Trans Multimedia*. 2013;15(5):983–1000.
65. Wein W. Brain-shift correction with image-based registration and landmark accuracy evaluation. In: *Simulation, image processing, and ultrasound systems for assisted diagnosis and navigation*. Cham: Springer; 2018. p. 146–51.
 66. Fuerst B, Wein W, Müller M, Navab N. Automatic ultrasound MRI registration for neurosurgery using the 2D and 3D LC2 metric. *Med Image Anal*. 2014;18(8):1312–9.
 67. Balakrishnan A, Zhao M, Sabuncu R, Guttad J, Dalca AV. VoxelMorph: a learning framework for deformable medical image registration. *IEEE Trans Med Imaging*. 2019;38(8):1788–800.
 68. Hu Y, et al. Weakly-supervised convolutional neural networks for multimodal image registration. *Med Image Anal*. 2018;49:1–13.
 69. Esteban J, Grimm M, Unberath M, Zahnd G, Navab N. Towards fully automatic X-ray to CT registration. In: *Medical image computing and computer assisted intervention – MICCAI*. Cham: Springer; 2019.
 70. Hou B, et al. Predicting slice-to-volume transformation in presence of arbitrary subject motion. In: *Medical image computing and computer assisted intervention – MICCAI*. Cham: Springer; 2017. p. 296–304.
 71. Bier B, et al. X-ray-transform invariant anatomical landmark detection for pelvic trauma surgery. In: *Medical image computing and computer assisted intervention – MICCAI*. Cham: Springer; 2018. p. 55–63.
 72. Gao C, Unberath M, Taylor R, Armand M. Localizing dexterous surgical tools in X-ray for image-based navigation. In: *Proceedings of IPCAI*. Cham: Springer; 2019. p. 1–4.
 73. Liao H, Lin W-A, Zhang J, Zhang J, Luo J, Zhou SK. Multiview 2D/3D rigid registration via a point-of-interest network for tracking and triangulation. In: *Proceedings of IEEE conference on computer vision and pattern recognition*. IEEE Computer Society; 2019. p. 12638–47.
 74. Gregory TM, Gregory J, Sledge J, Allard R, Mir O. Surgery guided by mixed reality: presentation of a proof of concept. *Acta Orthop*. 2018;89(5):480–3.
 75. Pratt P, Ives M, Lawton G, Simmons J, Radev N, Spyropoulou L, Amiras D. Through the Hololens' looking glass: augmented reality for extremity reconstruction surgery using 3d vascular models with perforating vessels. *Eur Radiol Exp*. 2018;2(1):2.
 76. Bergonzi L, Colombo G, Redaelli D, Lorusso M. An augmented reality approach to visualize biomedical images. *Comput Aided Des Appl*. 2019;16(6):1195–208.
 77. Sauer IM, Quiesner M, Tang P, Moosburner S, Hoepfner O, Horner R, Lohmann R, Pratschke J. Mixed reality in visceral surgery: development of a suitable workflow and evaluation of intraoperative use-cases. *Ann Surg*. 2017;266(5):706–12.
 78. Incekara F, Smits M, Dirven C, Vincent A. Clinical feasibility of a wearable mixed-reality device in neurosurgery. *World Neurosurg*. 2018;118:e422–7.
 79. Cartucho J, Shapira D, Ashrafian H, et al. Multimodal mixed reality visualisation for intraoperative surgical guidance. *Int J CARS*. 2020;15:819–26.
 80. Sinkin JC, Rahman OF, Nahabedian MY. Google glass in the operating room: the plastic surgeon perspective. *Plast Reconstr Surg*. 2016;138(1):298–302.
 81. Billings S, Deshmukh N, Kang HJ, Taylor R, Boctor EM. System for robot-assisted real-time laparoscopic ultrasound elastography. In: *SPIE medical imaging*. International Society for Optics and Photonics; 2012.
 82. Ruszkowski A, Moharer O, Lichtenstein S, Cook R, Salcudean S. On the feasibility of heart motion compensation on the DaVinci® surgical robot for coronary artery bypass surgery: implementation and user studies. In: *Robotics and automation (ICRA), 2015 IEEE international conference on*. IEEE; 2015. p. 4432–9.
 83. Pratt P, Hughes-Hallett A, Zhang L, Patel N, Mayer E, Darzi A, Yang G-Z. Autonomous ultrasound-guided tissue dissection. In: *Medical image computing and computer-assisted intervention – MICCAI 2015*. Springer; 2015.
 84. Hu D, Gong Y, Hannaford B, Seibel EJ. Semi-autonomous simulated brain tumor ablation with Ravenii surgical robot using behaviour tree. In: *Robotics and automation (ICRA), 2015 IEEE international conference on*. IEEE; 2015. p. 3868–75.
 85. Caversaccio M, Wimmer W, Anso J, Mantokoudis G, Gerber N, Rathgeb C, Schneider D, Hermann J, Wagner F, Scheidegger O, et al. Robotic middle ear access for cochlear implantation: first in man. *PLoS One*. 2019;14(8):e0220543.
 86. Zhang L, Ye M, Giataganas P, Hughes M, Yang G-Z. Autonomous scanning for endomicroscopic mosaicing and 3d fusion. In: *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE; 2017. p. 3587–93.
 87. Zhang L, Ye M, Giannarou S, Pratt P, Yang G-Z. Motion-compensated autonomous scanning for tumour localisation using intraoperative ultrasound. In: *International conference on medical image computing and computer-assisted intervention*. Springer; 2017. p. 619–27.
 88. Zhan J, Cartucho J, Giannarou S. Autonomous tissue scanning under free-form motion for intraoperative tissue characterisation. In: *2020 IEEE international conference on robotics and automation (ICRA)*. Paris: IEEE; 2020. p. 11147–54.
 89. Zhang L, Ye M, Giataganas P, Hughes M, Bradu A, Podoleanu A, et al. From macro to micro: autonomous multiscale image fusion for robotic surgery. *IEEE Robot Automat Mag*. 2017;24(2):63–72.
 90. Varghese RJ, Berthet-Rayne P, Giataganas P, Vitiello V, Yang G-Z. A framework for sensorless and autonomous probe-tissue contact management in robotic endomicroscopic scanning. In: *2017*

- IEEE international conference on robotics and automation (ICRA). IEEE; 2017. p. 1738–45.
91. Triantafyllou P, Wisanuvej P, Giannarou S, Liu J, Yang G-Z. A framework for sensorless tissue motion tracking in robotic endomicroscopy scanning. In: 2018 IEEE international conference on robotics and automation (ICRA). IEEE; 2018. p. 2694–9.
 92. Rosa B, Erden MS, Vercauteren T, Herman B, Szewczyk J, Morel G. Building large mosaics of confocal endomicroscopic images using visual servoing. *IEEE Trans Biomed Eng.* 2012;60(4):1041–9.
 93. Giataganas P, Hughes M, Payne CJ, Wisanuvej P, Temelkuran B, Yang G-Z. Intraoperative robotic-assisted large-area high-speed microscopic imaging and intervention. *IEEE Trans Biomed Eng.* 2018;66(1):208–16.
 94. O. Zettning, B. Frisch, S. Virga, M. Esposito, A. Rienmüller, B. Meyer, C. Hennersperger, Y.-M. Ryang, and N. Navab, “3d ultrasound registration-based visual servoing for neurosurgical navigation,” *Int J Comput Assist Radiol Surg*, vol. 12, no. 9, pp. 1607–1619, 2017.
 95. Virga S, Zettning O, Esposito M, Pfister K, Frisch B, Neff T, Navab N, Hennersperger C. Automatic force-compliant robotic ultrasound screening of abdominal aortic aneurysms. In: 2016 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE; 2016. p. 508–13.
 96. Merouche S, Allard L, Montagnon E, Soulez G, Bigras P, Cloutier G. A robotic ultrasound scanner for automatic vessel tracking and three-dimensional reconstruction of B-mode images. *IEEE Trans Ultrason Ferroelectr Freq Control.* 2015;63(1):35–46.
 97. Nadeau C, Krupa A, Petr J, Barillot C. Moments-based ultrasound visual servoing: from a mono-to multiplane approach. *IEEE Trans Robot.* 2016;32(6):1558–64.
 98. Pratt P, Hughes-Hallett A, Zhang L, Patel N, Mayer E, Darzi A, Yang G-Z. Autonomous ultrasound-guided tissue dissection. In: International conference on medical image computing and computer assisted intervention. Springer; 2015. p. 249–57.
 99. Chevrie J, Krupa A, Babel M. Real-time teleoperation of flexible beveled-tip needle insertion using haptic force feedback and 3d ultrasound guidance. In: 2019 international conference on robotics and automation (ICRA). 2019. p. 2700–6.
 100. Huang Q, Lan J, Li X. Robotic arm based automatic ultrasound scanning for three-dimensional imaging. *IEEE Trans Ind Inf.* 2018;15(2):1173–82.
 101. Hennersperger C, Fuerst B, Virga S, Zettning O, Frisch B, Neff T, Navab N. Towards mri-based autonomous robotic us acquisitions: a first feasibility study. *IEEE Trans Med Imaging.* 2016;36(2):538–48.
 102. Charalampaki P, Javed M, Daali S, Heiroth HJ, Igressa A, Weber F. Confocal laser endomicroscopy for real-time histomorphological diagnosis: our clinical experience with 150 brain and spinal tumor cases. *Neurosurgery.* 2015;62:171–6.
 103. Tzafetas M, Mitra A, Paraskevaidi M, Bodai Z, Kalliala I, Bowden S, Lathouras K, Rosini F, Szasz M, Savage A, Manoli E, Balog J, McKenzie J, Lyons D, Bennett P, MacIntyre D, Ghaem-Maghami S, Takats Z, Kyrgiou M. The intelligent knife (iKnife) and its intraoperative diagnostic advantage for the treatment of cervical disease. *Proc Natl Acad Sci USA.* 2020;117(13):7338–46.
 104. Desroches J, Jermyn M, Pinto M, Picot F, Tremblay M-A, Obaid S, Marple E, Urmy K, Trudel D, Soulez G, Guiot M-C, Wilson BC, Petrecca K, Leblond F. A new method using Raman spectroscopy for in vivo targeted brain cancer tissue biopsy. *Sci Rep.* 2018;8(1):1792.
 105. Ortega S, Fabelo H, Camacho R, De la Luz Plaza M, Callicó GM, Sarmiento R. Detecting brain tumor in pathological slides using hyperspectral imaging. *Biomed Opt Express.* 2018;9(2):818–31.
 106. André B, Vercauteren T, Buchner AM, Wallace MB, Ayache N. A smart atlas for endomicroscopy using automated video retrieval. *Med Image Anal.* 2011;15(4):460–76.
 107. André B, Vercauteren T, Buchner AM, Krishna M, Ayache N, Wallace MB. Software for automated classification of probebased confocal laser endomicroscopy videos of colorectal polyps. *World J Gastroenterol.* 2012;18(39):5560–9.
 108. André B, Vercauteren T, Buchner AM, Wallace MB, Ayache N. Learning semantic and visual similarity for endomicroscopy video retrieval. *IEEE Trans Med Imaging.* 2012;31(6):1276–88.
 109. Wan S, Sun S, Bhattacharya S, Kluckner S, Gigler A, Simon E, Fleischer M, Charalampaki P, Chen T, Kamen A. Towards an efficient computational framework for guiding surgical resection through intraoperative endo-microscopic pathology. In: Medical image computing and computer-assisted intervention (MICCAI). 2015. p. 421–9.
 110. Kamen A, Sun S, Wan S, Kluckner S, Chen T, Gigler AM, Simon E, Fleischer M, Javed M, Daali S, Igressa A, Charalampaki P. Automatic tissue differentiation based on confocal endomicroscopic images for intraoperative guidance in neurosurgery. *Biomed Res Int.* 2016;2016:6183218.
 111. Gu Y, Yang J, Yang GZ. Multi-view multi-modal feature embedding for endomicroscopy mosaic classification. In: 2016 IEEE conference on computer vision and pattern recognition workshops (CVPRW). 2016. p. 1315–23.
 112. Gu Y, Vyas K, Yang J, Yang GZ. Unsupervised feature learning for endomicroscopy image retrieval. In: Medical image computing and computer-assisted intervention (MICCAI). 2017. p. 64–71.
 113. Li Y, Charalampaki P, Liu Y, Yang GZ, Giannarou S. Context aware decision support in neurosurgical oncology based on an efficient classification of endomicroscopic data. *Int J Comput Assist Radiol Surg.* 2018;13(8):1187–1199. <https://doi.org/10.1007/>

- s11548-018-1806-7. Epub 2018 Jun 13. PMID: 29948845; PMCID: PMC6096753.
114. Ravi D, Szczotka AB, Pereira SP, Vercauteren T. Adversarial training with cycle consistency for unsupervised super-resolution in endomicroscopy. *Med Image Anal.* 2019;53:123–31.
115. Szczotka AB, Ravi D, Shakir DI, Pereira SP, Vercauteren T. Effective deep learning training for single-image super-resolution in endomicroscopy exploiting video-registration-based reconstruction. *Int J Comput Assist Radiol Surg.* 2018;13(6):917–24.
116. Baltussen EJM, Kok END, Brouwer de Koning SG, Sanders J, Aalbers AGJ, Kok NFM, Beets GL, Flohil CC, Bruin SC, Kuhlmann KFD, et al. Hyperspectral imaging for tissue classification, a way toward smart laparoscopic colorectal surgery. *J Biomed Opt.* 2019;24:016002.
117. Han Z, Zhang A, Wang X, Sun Z, Wang MD, Xie T. In vivo use of hyperspectral imaging to develop a non-contact endoscopic diagnosis support system for malignant colorectal tumors. *J Biomed Opt.* 2016;21:016001.
118. Pourreza-Shahri R, Saki F, Kehtarnavaz N, Leboulluec P, Liu H. Classification of ex-vivo breast cancer positive margins measured by hyperspectral imaging. In: Proceedings of the IEEE international conference on image processing, Melbourne, 15–18 September 2013, p. 1408–12.
119. Fei B, Lu G, Wang X, Zhang H, Little JV, Patel MR, Griffith CC, El-Diery MW, Chen AY. Label-free reflectance hyperspectral imaging for tumor margin assessment: a pilot study on surgical specimens of cancer patients. *J Biomed Opt.* 2017;22:086009.
120. Jayanthi JL, Nisha GU, Manju S, Philip EK, Jeemon P, Baiju KV, Beena VT, Subhash N. Diffuse reflectance spectroscopy: diagnostic accuracy of a non-invasive screening technique for early detection of malignant changes in the oral cavity. *BMJ Open.* 2011;1:e000071.
121. Regeling B, Laffers W, Gerstner AOHH, Westermann S, Müller NA, Schmidt K, Bendix J, Thies B. Development of an image pre-processor for operational hyperspectral laryngeal cancer detection. *J Biophotonics.* 2016;9:235–45.
122. Ravi D, Fabelo H, Callic GM, Yang GZ. Manifold embedding and semantic segmentation for intraoperative guidance with hyperspectral brain imaging. *IEEE Trans Med Imaging.* 2017;36: 1845–57.
123. Fabelo H, Ortega S, Ravi D, Kiran BR, Sosa C, Bulters D, Callicó GM, Bulstrode H, Szolna A, Piñeiro JF, et al. Spatio-spectral classification of hyperspectral images for brain cancer detection during surgical operations. *PLoS One.* 2018;13:e0193721.
124. Halicek M, Lu G, Little JV, Wang X, Patel M, Griffith CC, El-Deiry MW, Chen AY, Fei B. Deep convolutional neural networks for classifying head and neck cancer using hyperspectral imaging. *J Biomed Opt.* 2017;22:060503.
125. Halicek M, Little JV, Wang X, Chen AY, Fei B. Optical biopsy of head and neck cancer using hyperspectral imaging and convolutional neural networks. *J Biomed Opt.* 2019;24:036007.
126. Fabelo H, Halicek M, Ortega S, Szolna A, Morera J, Sarmiento R, Callicó GM, Fei B. Surgical aid visualization system for glioblastoma tumor identification based on deep learning and in-vivo hyperspectral images of human patients. In: Fei B, Linte CA, editors. *Medical imaging 2019: image-guided procedures, robotic interventions, and modeling*, vol. 10951. San Diego: International Society for Optics and Photonics; 2019. p. 35.
127. Fabelo H, Halicek M, Ortega S, Shahedi M, Szolna A, Piñeiro J, Sosa C, O'Shanahan A, Bisshop S, Espino C, et al. Deep learning-based framework for in vivo identification of glioblastoma tumor using hyperspectral images of human brain. *Sensors.* 2019;19:920.



Artificial Intelligence in Surgery

61

Filippo Filicori and Ozanan R. Meireles

Contents

Introduction	856
AI-Powered Techniques	856
Computer Vision	857
Natural Language Processing	858
Current Applications of AI in Surgery	858
Preoperative Risk Prediction	858
Intraoperative Video Analysis	859
Surgical Workflow Analysis	860
Regulatory and Legal Considerations	860
Conclusion	861
References	861

Abstract

Artificial intelligence (AI) is the study of algorithms that give machines the ability to reason and perform cognitive functions. Applications of AI in medicine broadly and surgery, more specifically, have grown over the last few years

as technology has advanced and clinical data has become more digitally accessible. It is becoming more important for surgeons to develop a fundamental understanding of the common techniques, applications, limitations, and ethical considerations of AI in surgery. This chapter provides an overview of AI for surgeons and describes ways in which surgeons can play a role in future development of AI applications.

F. Filicori

Intraoperative Performance Analytics Laboratory,
Department of Surgery, Lenox Hill Hospital, Hofstra
School of Medicine at Northwell, New York, NY, USA
e-mail: ffilicori@northwell.edu

O. R. Meireles (✉)

Surgical Artificial Intelligence and Innovation Laboratory,
Department of Surgery, Massachusetts General Hospital,
Boston, MA, USA
e-mail: ozmeireles@mgh.harvard.edu

Keywords

Artificial intelligence · Deep learning ·
Computer vision · Natural language
processing · Learning algorithms · Risk
prediction · Surgical automation · Robotic
surgery · Critical view of safety

Introduction

Artificial intelligence (AI) has been long applied to many medical fields for the purposes of diagnostics, decision-making, or clinical research. The application of such technology in surgery will bring unparalleled benefits such as enhanced patient selection/management, a truly automated operating room environment with intraoperative assistance to the surgeon, and a real-time surgical workflow detection which will result in increased efficiency and safety.

Much of the recent advances in the application of AI in surgery were recently fostered by computer vision (CV). Such application has brought the ability to provide in-depth analysis of intraoperative surgical footage with the short-term goal of providing live intraoperative feedback and the long-term goal of partial automation of certain surgical procedures. CV is, in its simplest explanation, machine understanding of images and videos [1]. It constitutes the way machines understand and interact with our reality. The first goal of such interaction will be plain understanding followed by cognitive augmentation during a surgical procedure. Augmentation will most likely take the form of intraoperative live feedback to the operator. Current efforts are focused on developing better CV algorithms which “understand” a surgical operation; however, such endeavor is a challenging task to achieve.

To provide a comparison with the automotive industry, the Society of Automotive Engineers has provided six possible levels of automation in cars ranging from 0 (no automation) to 5 (full automation and lack for requirement of oversight by the driver) [2]. Despite some early promises of reaching level 5 by the end of the decade, billions of dollars of investments, and almost half a century of work by the industry and academia, we are presently only able to produce vehicles with partial automation and that still require partial driver’s oversight (level 3 automation). Although comparisons are somewhat difficult to make, cognitive augmentation could be assimilated to level 1 automation, much like a car can provide notifications and even real-time intervention if you are

steering off your lane on the highway or if an obstacle is in front of you and you need to brake. In a similar fashion, you could be soon operating with a “Surgical GPS” which can tell you in real time whether you are about to cut the wrong anatomical structure or if you are likely to experience postoperative bleeding at the end of an operation.

Much of the hurdles that we have to face to provide live feedback to the surgeon hinge on data processing. The average self-driving car requires tens of thousands of sensors and processes an average of 100 terabyte of data every 8 h. A “Surgical GPS” will likely require a similar data handling capability. This brings unparalleled challenges both in terms of acquisition and processing of data.

For such reason, in the short run, current technology still focuses on early applications such as automated performance analysis in the postoperative setting. Currently, automated analysis of surgical videos was successfully used for video segmentation [3], identification of core anatomical features such as the critical view of safety during laparoscopic cholecystectomy [4], and determination of adequacy of a lymph node dissection [5]. The next logical step toward cognitive augmentation will likely require FDA clearance since such software will likely bring real-time modifications in the intraoperative decision-making.

AI-Powered Techniques

AI is a larger parent field that encompasses sub-fields like machine learning and computer vision, which further encompasses techniques like neural networks and deep learning. These subfields and others are actually quite interrelated, and techniques from one may be subsumed or used in concert with another, depending on the application.

Neural networks, inspired by biological nervous systems, process data in layers of simple computational units that are intended to be analogous to neurons. Thus, unlike in classical machine learning, neural networks can extract features

from data and use them as inputs, adjusting the weights of those features accordingly to be used within an activation function to yield some output, [6] that is, the system automatically, using predetermined mathematical functions, tweaks weights to strengthen/weaken connections within the network to yield the best possible results.

Deep neural networks are, effectively, a neural network with more than three layers, allowing for learning of more complex patterns than those that are discernible from simple one- or two-layer networks. As with “non-deep” neural networks, deep learning will select features that are most likely to yield best results. This technique works particularly well with unstructured data such as audio, images, and video. Generally, each layer of a deep neural network performs a set of operations to generate a representation of the data that is then fed to the next layer.

Computer Vision

Technical Aspects of Computer Vision in Surgery

Computer vision is a field of artificial intelligence that uses deep machine learning models in order for computers to understand and evaluate their surroundings. Computer vision and image processing technology are being developed in order to read, interpret, segment, analyze, and assess surgical videos. One goal of computer vision and robotic technology is to create a surgical robot that can analyze a surgical situation to make corrections intraoperatively, but the ultimate goal is to design a robot that can operate autonomously.

Computer Vision and Supervised Learning

Within machine learning (ML), the two most common learning types are supervised and unsupervised learning. Supervised learning is a task-driven process wherein an algorithm is trained to predict a prespecified output, such as identifying a stop sign or recognizing a cat in a photograph. The “supervised” term comes from the need to provide annotated (i.e., labeled) data

so that it can learn the associations between inputs and the desired output. Several software are available to provide the user with annotation capabilities. Annotating is a time-consuming endeavor where both temporal sequences are determined (e.g., determining the time stamp in the different phases of an operation) and anatomical structures are labeled. Thus, datasets are divided into a training set (with labels provided) for learning and a test set (no labels provided) that allows for the assessment of the performance of the algorithm on new data [1, 7]. Such application of ML is particularly useful in surgery given the current technological options. Recognition of anatomical structures starting from a known dataset is in fact one of the early goals of surgical automation and the foundation of intraoperative feedback. Such modality, however, if fraught by the particularly long-time consumption required as annotating surgical images frame by frame requires several hours of work from an experienced operator. Temporal annotations are similarly time-consuming. The purpose of those and the exact sequence determination during a surgery is, however, equally important. Applications of such annotation range from monitoring the workflow during a surgery and allowing for a more streamlined time management strategy to determining when a training surgeon is struggling or identifying more difficult portions of a surgical operation once you are reviewing the footage.

Computer Vision and Unsupervised Learning

Unlike supervised learning, unsupervised learning does not utilize a prespecified annotation; rather, it draws inferences from unlabeled data to identify patterns and/or structure within a dataset. This type of learning can be useful in identifying relationships between groups (e.g., clustering) for further hypothesis generation. This can be applied to typical, discrete surgical data such as patient outcome databases or to more unique datasets such as surgical motion and activity. For example, unsupervised learning has been used to identify high-risk cardiac surgery patients and to automatically identify suturing motion in surgical video

[8, 9]. A third category of learning is reinforcement learning, a form of unsupervised learning. It is analogous to operant conditioning, where learning occurs through successive attempts via trial and error and rewards/punishments guide the behavior of the model to optimize rewards. The most common architectures in deep learning that are currently being used for surgical applications are convolutional neural networks (CNNs), recurrent neural networks (RNNs), and residual neural networks (ResNets).

Natural Language Processing

Natural language processing (NLP) focuses on machine understanding of human language beyond identification of vocabulary (e.g., synonyms, antonyms, definitions, etc.). Without NLP, computers are limited to reading machine languages or code (e.g., C+, Java, Visual Basic) to execute instructions based on explicitly programmed codes that are compiled to yield an output. NLP allows machines to approximate the understanding of human language as it would be used in day-to-day life. It strives to achieve understanding of syntax and semantics to approximate meaning from phrases, sentences, or paragraphs [10].

NLP is perhaps most readily recognized in home assistant devices such as Amazon Alexa (Amazon, Seattle, WA) or Google Home (Alphabet, Mountain View, CA). Analogous functions are found in digital platforms used for operative dictation (e.g., Nuance's Dragon software [Nuance, Burlington, MA]). Beyond the provider-facing functions such as dictation, NLP is utilized heavily in the analysis and utilization of data within the electronic medical record. Because NLP can be used to analyze some forms of human language, unstructured free text such as radiology reports, progress reports, and operative notes can be analyzed and structured in an automated manner. As examples, it can be utilized to assess for sentiment in patient notes for the prediction of patient health status, to analyze records for risk prediction in cancer patients, or to detect surgical site infection from providers' notes [10–12].

Current Applications of AI in Surgery

Preoperative Risk Prediction

As up to 20% of surgical procedures can result in a complication [13], preoperative risk stratification is of the utmost importance for therapeutic planning. Ideally, risk prediction would both guide patient-centered decisions to evaluate operative candidacy and predict possible postoperative complications.

Many risk calculators and decision algorithms exist on the market. Since adequate cardiac function is of utmost importance, the most prominent calculators assess and predict risk of major adverse cardiac events (MACE). Examples include the Revised Cardiac Risk Index (RCRI) and Gupta Perioperative Risk for Myocardial Infarction or Cardiac Arrest (MICA) [14, 15]. Unfortunately, these models often underperform. The POISE trials showed that a MACE rate of 6.9% in patients rather than the RCRI-predicted rate of 1–2.4%. Outside of risk calculators, subjective patient-reported measures of cardiac functional capacity, such as metabolic equivalents (METs), also tend to under-triage patients [16]. One study found that subjects report their cardiac capacity with only a sensitivity of 19% [17].

More recent efforts attempt to solve the prior predictors' shortcomings through the use of objective "big data" to address model underperformance. For example, the American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) released a risk calculator. Their calculator used information from 393 hospitals with almost 1.5 million patients to create a generalized linear mixed model to predict risk of mortality and various complications. This model had good performance, with a c-statistic of 0.944 and 0.816 for mortality and morbidity, respectively [18].

Until recently, the majority of risk calculators have used traditional linear and additive models for risk prediction. Recent advances utilize machine learning (ML) methods for better approximation of the non-linearity of patient risk factors. Researchers at Duke University trialed

3 different ML methods on their single-institution database of 100,000 patients: least absolute shrinkage and selection operator (LASSO) penalized logistic regression, random forest models, and extreme gradient boosted trees [19]. Comparing the three methods, they found that LASSO performed best for over 8/14 outcomes, while extreme gradient boosted trees excelled in 5/14 outcomes with area under the curve (AUC) ranging from 0.747 to 0.924. With their algorithms, they created an online calculator with 9 input data fields that outperformed the ACS-NSQIP calculator in postoperative mortality and morbidity prediction across the board for a random sample of 75 patients.

Similar work has come out of the University of Florida with their MySurgeryRisk score. They used their EMR data to create risk prediction scores using ML techniques such as random forests. Interestingly, their risk prediction was particularly patient-tailored since they linked training data to census data tied to ZIP codes and to surgeon-specific outcomes. Beyond just the creation of a risk calculator, they also created interfaces for seamless EMR integration, so that not only risk prediction happened in real time but their models underwent continuous learning and tuning from physician feedback [20].

Preoperative risk calculation has evolved over the past decades. Medical practitioners can combine the ever-increasing big data from EMR with ML algorithms for objective, and increasingly accurate, predictions of patient outcomes. With continuous EMR integration and even deployment to smartphones, this wealth of information is immediately available and creates the promise that one day, we may be able to exactly answer our patients when they ask “What are the risks of this surgery?”

Intraoperative Video Analysis

Most of the readily recognizable advances in CV have come from the fields of radiology and pathology, perhaps due to the readily available nature of digital images in both fields. CV has also demonstrated promise with screening applications in

ophthalmology, such as automated detection of diabetic retinopathy, and dermatology, where automated recognition of benign versus malignant skin lesions has been described [21, 22].

However, CV applications in surgery are increasing as access to intraoperative footage increases. With greater, cheaper storage capacities and more user-friendly laparoscopic, endoscopic, and robotic camera systems, many surgeons are choosing to record their operations for teaching, education, and research purposes.

AI technology, through CV, allows computers to comprehend visual cues and therefore interact with the world in real time. With sufficient training and incorporation of thousands of operations, an AI model could guide surgeons, in real time, just as if they had the world expert in surgery looking over their shoulder. We already know that experience matters, with an inverse relationship between a surgeon’s case volume and their patient’s mortality [23]. Analysis of a surgeon’s ability on visual cues alone is even predictive of a surgeon’s rate of complications [24]. If a CV system could guide surgeons and take their performance from the bottom to the top quartile, patients would receive immediate improvement in their care. We know that almost 70% of cases have “near-miss” events of which two-thirds will need additional intervention to fix, perhaps something a CV model could warn the operator ahead of time and prevent from happening in the first place [25].

With the promise of a safer OR, a few groups across the world have tried to tackle the difficult problem of teaching a computer to see and think like a high-level surgeon. CV is still quite new, with accurate image recognition only possible since 2012; therefore, applications of CV in surgery are still in their early stages [25, 26]. The initial work has involved analysis of laparoscopic cases given the ease of video acquisition and reproducibility of surgical images. In particular, groups have worked on identification of surgical phases of an operation with good accuracy across cholecystectomy (86.7%), sleeve gastrectomy (85.6%), and sigmoidectomy (91.9%) [3, 27, 28]. Additional applications of such technology have been investigated for its potential impact on

improving operating room workflow and logistics such as through the prediction of remaining operative time from intraoperative video alone [29].

Knowing that accurate phase recognition is possible, the next steps will include development of intraoperative decision support. For example, likely applications include guidance for port placement, confirmation that a critical portion of the case has successfully been obtained (e.g., the critical view of safety in cholecystectomy [38] or adequate dissection during an inguinal hernia), and, in the more distant future, even a real-time intraoperative “GPS” to guide surgeons in their dissection. As the AI models train with an increasing number of cases, they will soon develop an unparalleled surgical knowledge – a “collective surgical consciousness” – that will help any surgeon, anywhere, to deliver optimal intraoperative care to their patients [30]. Similar to a chess game, a player has to play so many games and take so many different approaches to become proficient, but he can only learn one game at a time. If all the chess players in the world could share their collective knowledge with each other through AI, the learning would be exponentially higher and faster. This is the concept of a collective consciousness. By using AI, surgeons could be sharing their individual experiences to generate a similar collective consciousness for surgery.

There are, however, key advances that must be made. An important, early advance in the process of translating computer vision to the operating room is the establishment of clear labels for operative videos. Hashimoto et al. demonstrated that surgeons, even within the same institution, can differ in their conceptualization of the boundaries of the steps of an operation, [3] that is, when does one step of an operation end and the next begin? As previously described, for supervised learning, defining a “gold standard” or ground truth is important to be able to train a model to recognize aspects of surgical video [4]. Establishing ground truth can be difficult when a surgical operation has several different variants and there might no single correct way to perform it. Efforts are currently underway through the Society of American Gastrointestinal and Endoscopic Surgeons (SAGES)

to convene an international consensus on guidelines for annotating operative video for the purposes of machine learning and computer vision research.

Surgical Workflow Analysis

Several applications were recently devised to both increase surgical safety and enhance performance in the operating room [31]. Most of them rely on the identification of the personnel in the operating room through environmental cameras. Identification is done automatically through visual cues that are captured by the cameras through CV. As such, the software can automatically detect the surgical workflow in the operating room. Common events which can be captured include the patient entering the operating room, the scrub tech setting up the instrument tray, or the circulating nurses holding up a brightly colored board during surgical timeout. Automatic workflow detection has the potential to enhance both productivity and safety of the surgical environment [32]. As more complex software is currently being devised to recognize motion of the surgeons in greater detail, this will invariably increase the yield of in-depth analysis of the operating room environment in real time.

Regulatory and Legal Considerations

AI applications in surgery are a growing reality. Surgeons should familiarize themselves with this technology and its medicolegal implications, including intellectual property and data ownership, consent and scope of videos, privacy, and the potential for litigation. Highlighting the specific nature of these algorithms, the Food and Drug Administration (FDA) approved the first diagnostic utilization of an AI algorithm in 2018 – a program that assists in screening for diabetic retinopathy through automated analysis of images of the fundus [2]. The list of FDA-approved algorithms continues to grow with approved applications in radiology, cardiology, and pathology as well. With ongoing

development and application of AI technologies in medicine, it is important for clinicians in every field to understand what these technologies are and how they can be leveraged to deliver safer, more efficient, more cost-effective care. Furthermore, it is important to keep in mind that these technologies are not a panacea and to understand the limitations inherent to any tool.

Conclusion

In summary, artificial intelligence as applied to surgery is early in its development. While significant advances are being made in AI, these advances are focused on narrow applications of the technology to specific problems within surgery. The field is very much in a phase of discovery and development, and a critical appraisal of new publications, software, and devices is necessary to appropriately evaluate its impact on patient care and surgeon workflow. As with any new technology, a healthy measure of skepticism is necessary to guard against hype; however, data on potential applications of AI to surgery have thus far been promising.

References

- Bellman R. An introduction to artificial intelligence: can computers think? Thomson Course Technology; 1978.
- Administration USF & D, U.S. Food & Drug Administration. FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems. Case Med Res. 2018. <https://doi.org/10.31525/fda2-ucm604357.htm>.
- Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial intelligence in surgery: promises and perils. Ann Surg. 2018;268(1):70–6.
- Hashimoto DA, Witkowski E, Gao L, Meireles O, Rosman G. Artificial intelligence in anesthesiology. Anesthesiology. 2019;132:379. <https://doi.org/10.1097/alan.0000000000002960>.
- Avanzolini G, Barbini P, Gnudi G. Unsupervised learning and discriminant analysis applied to identification of high risk postoperative cardiac patients. Int J Biomed Comput. 1990;25(2–3):207–21. [https://doi.org/10.1016/0020-7101\(90\)90010-r](https://doi.org/10.1016/0020-7101(90)90010-r).
- DiPietro R, Hager GD. Unsupervised learning for surgical motion by learning to predict the future, vol. 2018. Medical Image Computing and Computer Assisted Intervention – MICCAI; 2018. p. 281–8. https://doi.org/10.1007/978-3-030-00937-3_33.
- Skinner BF. The behavior of organisms: an experimental analysis. B. F. Skinner Foundation; 1990.
- Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of go without human knowledge. Nature. 2017;550(7676):354–9.
- Hebb DO. The organization of behavior. Taylor and Francis; 2005. <https://doi.org/10.4324/9781410612403>.
- Natarajan P, Frenzel JC, Smaltz DH. Demystifying big data and machine learning for healthcare. Taylor and Francis; 2017. <https://doi.org/10.1201/9781315389325>.
- Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24–9.
- Grzybowski A, Brona P, Lim G, et al. Artificial intelligence for diabetic retinopathy screening: a review. Eye. 2019. <https://doi.org/10.1038/s41433-019-0566-0>.
- Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–8.
- Rodas NL, Padov N. Augmented reality for reducing intraoperative radiation exposure to patients and clinicians during x-ray guided procedures. In: Mixed and augmented reality in medicine. CRC Press; 2018. p. 217–29. <https://doi.org/10.1201/9781315157702-15>.
- Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc. 2011;18(5):544–51.
- Hughes KS, Zhou J, Bao Y, Singh P, Wang J, Yin K. Natural language processing to facilitate breast cancer research and management. Breast J. 2020;26(1):92–9.
- Zunic A, Corcoran P, Spasic I. Sentiment analysis in health and well-being: systematic review. JMIR Med Inform. 2020;8(1):e16023.
- Shen F, Larson DW, Naessens JM, Habermann EB, Liu H, Sohn S. Detection of surgical site infection utilizing automated feature generation in clinical notes. Int J Healthc Inf Syst Inform. 2019;3(3):267–82.
- Healey MA, Shackford SR, Osler TM, Rogers FB, Burns E. Complications in surgical patients. Arch Surg. 2002;137(5):611–7; discussion 617–618.
- Lee TH, Marcantonio ER, Mangione CM, et al. Derivation and prospective validation of a simple index for prediction of cardiac risk of major noncardiac surgery. Circulation. 1999;100(10):1043–9.
- Gupta PK, Gupta H, Sundaram A, et al. Development and validation of a risk calculator for prediction of cardiac risk after surgery. Circulation. 2011;124(4):381–7.
- POISE Study Group, Devereaux PJ, Yang H, et al. Effects of extended-release metoprolol succinate in patients undergoing non-cardiac surgery (POISE trial): a randomised controlled trial. Lancet. 2008;371(9627):1839–47.

23. Wijeysundera DN, Pearse RM, Shulman MA, et al. Assessment of functional capacity before major non-cardiac surgery: an international, prospective cohort study. *Lancet.* 2018;391(10140):2631–40.
24. Wolters U, Wolf T, Stützer H, Schröder T. ASA classification and perioperative variables as predictors of postoperative outcome. *Br J Anaesth.* 1996;77(2):217–22.
25. Owens WD, Felts JA, Spitznagel EL Jr. ASA physical status classifications: a study of consistency of ratings. *Anesthesiology.* 1978;49(4):239–43.
26. Bilimoria KY, Liu Y, Paruch JL, et al. Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. *J Am Coll Surg.* 2013;217(5):833–42.e1–e3.
27. Corey KM, Kashyap S, Lorenzi E, et al. Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): a retrospective, single-site study. *PLoS Med.* 2018;15(11):e1002701.
28. Bihorac A, Ozrazgat-Baslanlı T, Ebadi A, et al. MySurgeryRisk: development and validation of a machine-learning risk algorithm for major complications and death after surgery. *Ann Surg.* 2019;269(4):652–62.
29. Bertsimas D, Dunn J, Velmahos GC, Kaafarani HMA. Surgical risk is not linear. *Ann Surg.* 2018;268(4):574–83. <https://doi.org/10.1097/SLA.000000000002956>.
30. Birkmeyer JD, Stukel TA, Siewers AE, Goodney PP, Wennberg DE, Lucas FL. Surgeon volume and operative mortality in the United States. *ACC Curr J Rev.* 2004;13(2):59. <https://doi.org/10.1016/j.accreview.2003.12.065>.
31. Birkmeyer JD, Finks JF, O'Reilly A, et al. Surgical skill and complication rates after bariatric surgery. *N Engl J Med.* 2013;369(15):1434–42.
32. Bonrath EM, Gordon LE, Grantcharov TP. Characterising “near miss” events in complex laparoscopic surgery through video analysis. *BMJ Qual Saf.* 2015;24(8):516–21. <https://doi.org/10.1136/bmjqqs-2014-003816>.
33. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM.* 2017;60(6):84–90. <https://doi.org/10.1145/3065386>.
34. Hashimoto DA, Rosman G, Witkowski ER, et al. Computer vision analysis of intraoperative video: automated recognition of operative steps in laparoscopic sleeve gastrectomy. *Ann Surg.* 2019;270(3):414–21.
35. Kannan S, Yengera G, Mutter D, Marescaux J, Padov N. Future-state predicting LSTM for early surgery type recognition. *IEEE Trans Med Imaging.* 2019. <https://doi.org/10.1109/TMI.2019.2931158>.
36. Kitaguchi D, Takeshita N, Matsuzaki H, et al. Real-time automatic surgical phase recognition in laparoscopic sigmoidectomy using the convolutional neural network-based deep learning approach. *Surg Endosc.* 2019. <https://doi.org/10.1007/s00464-019-07281-0>.
37. Twinanda AP, Yengera G, Mutter D, Marescaux J, Padov N. RSDNet: learning to predict remaining surgery duration from laparoscopic videos without manual annotations. *IEEE Trans Med Imaging.* 2019;38(4):1069–78.
38. Mascagni P, Fiorillo C, Urade T, et al. Formalizing video documentation of the Critical View of Safety in laparoscopic cholecystectomy: a step towards artificial intelligence assistance to improve surgical safety. *Surg Endosc.* 2019. <https://doi.org/10.1007/s00464-019-07149-3>.



Artificial Intelligence in Urology

62

Kevin Y. Chu and Michael B. Tradewell

Contents

Introduction	864
Artificial Intelligence in Urology	864
Urologic Oncology	864
Endourology	867
Andrology	868
Conclusion	870
References	870

Abstract

Urology is intertwined with surgical technology. Endoscopic and robotic surgery found early acceptance within the practice of urology. Artificial intelligence research has begun to permeate through the urologic literature. In urologic oncology, AI systems have been developed to diagnose malignancy, guide therapeutics, and predict surgical outcomes. These systems have been shown to grade prostate cancer biopsies more accurately than general pathologists and can accurately predict postoperative length of stay based upon robotic laparoscope kinetics. AI enables accurate

prediction of kidney stone passage and stone clearance rates after surgery. Robotic systems using AI have successfully guided renal puncture in early clinical trials. The evaluation and treatment of the infertile male is seeing a paradigm shift as AI systems predict fertility potential and sperm retrieval success. In the future, AI algorithms may inform sperm retrieval for in vitro fertilization optimization. While many of the aforementioned AI systems remain isolated within single-institution research endeavors, published online AI predictors make these analyses accessible to general urologists. In time, AI can be expected to take a larger foothold into modern urologic practice.

Keywords

Artificial intelligence · Urology · Machine learning · Urologic oncology · Endourology ·

K. Y. Chu · M. B. Tradewell (✉)
Department of Urology, University of Miami Miller
School of Medicine, Miami, FL, USA

Infertility · Predictive modeling · Surgical robotics

Introduction

Urology has always been at the forefront of research and innovation within the medical community. The field was one of the early adopters of endoscopic technology by implementing cystoscopy into the urological workflow. Endoscopy quickly became an integral diagnostic and therapeutic tool for a wide variety of urological ailments. Urologists also established one of the first biomarkers for cancer surveillance in prostate-specific antigen. While there are many other notable examples of these advancements in urology, it was the early adoption of robotic surgery into practice that reflected the field's propensity for innovation [1]. Artificial intelligence (AI) is the next technological frontier with promise for significant advancements in patient care, and its application has begun permeating the urologic literature.

Artificial Intelligence in Urology

Urologic Oncology

History

Urologic oncology has a rich history of implementing predictive models to guide clinical practice. Notably, the 1993 Partin Tables used a nomogram with inputs of preoperative prostate-specific antigen (PSA) level, clinical stage, and Gleason score on biopsy to predict pathological stage at the time of radical prostatectomy [2]. These analyses remain the standard in predicting pathological stage and subsequent risk stratification of men newly diagnosed with prostate cancer. Similarly, urologic oncologists are well accustomed to adoption of new technologies. While urologists were not the first to use the Da Vinci Robotic Surgical System (Intuitive, Sunnyvale, California), the widespread use of the technology for robotic-assisted laparoscopic radical prostatectomy was critical for the adoption of the

robotic systems into modern-day surgery over the last two decades [3]. Not surprisingly, the use of artificial intelligence (AI) has made some early inroads into the modern practice of urologic oncology.

Prostate Cancer

Prostate cancer is the area of urology with the most robust application of AI to date. AI has been used to determine Gleason grades on prostate biopsy and MRI, predict postoperative surgical outcomes, and guide optimal treatment regimens. Auffenberg et al. used a state-wide database from Michigan to create a random forest machine learning model to predict treatment decisions based on patients with similar characteristics. Model inputs included pre-treatment data from 7543 men diagnosed with prostate cancer for 45 urology practices across the state. Outputs were primary treatment prediction including radical prostatectomy, radiation therapy (either external beam, brachytherapy, or both), primary androgen deprivation therapy, active surveillance, and watchful waiting. Using a 2:1 train-test split, the random forest performed with an AUC of 0.82 and good calibration [4]. These data may be useful to inform critical shared decisions made during localized prostate cancer care. Other AI models have been implemented to improve prostate cancer diagnostics.

Nagpal et al. developed a deep learning algorithm to predict Gleason grade group from prostate biopsy specimens [5]. These analyses included 752 prostate needle core biopsy specimens from 4 institutions. Ground-truth grade group was determined by concordance from two expert pathologists. The deep learning algorithm was trained and validated on 752 biopsy specimens including 322 external validation specimens from a single institution not used in algorithm training and tuning. Model performance was assessed by comparing majority opinions from 19 general pathologists. The deep learning algorithm grade group determination agreed more often with expert pathologists (71.7%; 95% CI, 67.9–75.3%) than general pathologists did (58.0%; 95% CI, 54.5%–61.4%) ($P < 0.001$). The authors concluded the algorithm showed

higher proficiency than general pathologists at Gleason grading prostate needle core biopsy specimens on an external validation data set.

Subsequent analysis by Steiner et al. further assessed the aforementioned deep learning algorithm as a decision support tool to improve the accuracy of prostate biopsy Gleason grade grouping [6]. Twenty general pathologists reviewed 240 prostate biopsy specimens with ground truth established by concordance from 2 expert genitourinary pathologists. Pathologists were randomized into two groups reviewing cases in batches of ten alternating between AI assistance and without. After a 4-week washout period, the pathologists reviewed the same cases with the opposite assistance or without. The author's found artificial intelligence-assisted review caused a 5.6% increase (95% CI, 3.2%–7.9%; $P <0.001$) in agreement with specialists (from 69.7% for unassisted reviews to 75.3% for assisted reviews). AI assistance decreased review time by the pathologist 13.5%. These data show the promise of increasing diagnostic accuracy and efficiency with artificial intelligence. However, the algorithm could not achieve the levels of expert genitourinary pathologist and could not elevate the performance of a general pathologist to that of an expert.

In the past decade, magnetic resonance imaging (MRI) has become an integral tool in urology for identifying potential prostatic lesions of clinical significance that may have been missed on standard template systematic transrectal ultrasound-guided prostate biopsies. With the advent of technology to allow for MRI-ultrasound fusion biopsy of the prostate, patients are provided more targeted diagnostics and results. Current classification of prostatic lesions is according to the Prostate Imaging-Reporting and Data System (PIRADS) criteria. While there is a consensus on lesion features for classification, there remains a subjective bias in grading. Schelb et al. utilized a deep learning algorithm (U-Net) to combine T2-weighted and diffusion MR imaging phases and found potential in supporting clinical classification of prostatic lesions. Sensitivity to detect clinically significant prostate cancer was found to be 88% versus 92%

(PIRADS vs. U-Net, $p >0.99$) and specificity 50% versus 47% ($p >0.99$). Collating data between U-Net and clinical radiologist assessment resulted in improved positive predictive value for clinically significant prostate cancer from 48% to 67% ($p = 0.01$), and negative predictive value remained unchanged ($p >0.99$) [7]. Additionally, MRI may contain texture features of the prostatic lesion that reveal further information regarding the malignant pathology yet are indistinguishable to the human eye. Fehr et al. combined apparent diffusion coefficient (ADC) and T2-weighted MR imaging phases with sample augmentation and utilized recursive feature selection support vector machine (RFE-SVM) to robustly predict (a) cancerous vs. noncancerous lesions, (b) low- vs. intermediate-/high-risk prostate cancer, and (c) sub-classification of intermediate-risk prostate cancer [8]. Their work suggests the feasibility to classify prostate cancer with imaging despite highly imbalanced data. As AI applications continue to evolve in urologic radiology for prostate cancer, these will improve diagnostic information available to patients and physicians for shared decision-making.

Urologists and computer scientists from the Center for Robotic Simulation and Education at the USC Keck School of Medicine have implemented AI to predict postoperative outcomes based on Da Vinci Surgical System kinematic data during robot-assisted laparoscopic radical prostatectomy. A random forest model was trained to predict postoperative length of stay (≤ 2 days or > 2 days) from 25 various intraoperative instrument, camera, and energy usage metrics in 78 cases from 9 surgeons (67 with length of stay ≤ 2 days and 11 > 2 days). The model performed with an accuracy of 87.5% accessed with k-fold ($k = 10$) cross-validation. Features related to camera manipulation were most predictive [9]. In a subsequent analysis, the research group developed a deep learning algorithm to predict urinary continence after robot-assisted laparoscopic radical prostatectomy from Da Vinci kinematics data. Using 100 cases from 8 surgeons, a deep learning model performed with an moderate predictive

accuracy measured by a mean-absolute error on 85.9 days and a cindex of 0.6. While accurate prediction of incontinence has a significant room for improvement, the three most informative inputs were the kinematics of suturing the anastomosis between the bladder and urethra [10]. Most importantly, these papers shed light on the potential AI applications in surgical performance evaluation.

Kidney Cancer

Recent clinical advances in AI-applied urological oncology extend beyond prostate cancer. Renal cancer has been well represented in recent years at the Medical Image Computing and Computer-Assisted Intervention (MICCAI) conference grand challenges. These grand challenges serve to elevate medical image algorithms through open data and outcome sharing. The KiTS19 challenge saw 106 teams from 5 continents compete to develop an AI algorithm capable of auto-segmenting kidneys and kidney tumors from CT images. The challenge consisted of 300 CT data sets with segmentation masks. Teams trained on 210 paired data sets and tested their algorithms on 90 CT images producing the predicted segmentation masks. The winning team achieved an average Sorenson-Dice coefficient of 0.974 for kidney and 0.851 for tumor. These data approached the inter-annotator Dice for kidney segmentation (0.983) but not for tumor (0.923) [11]. A similar EndoVis sub-challenge 2017 Kidney Boundary Detection used laparoscopic nephrectomy video files with labeled pixels containing kidney edges to create algorithms to automatically segment kidney boundaries [12]. A future KiTS21 MICCAI grand challenge will expand on the 2019 challenge to include segmentation of other anatomical features relevant to nephrectomy and partial nephrectomy, including the ureter and renal vasculature.

For a successful partial nephrectomy, a surgeon must successfully identify anatomical features and the tumor itself while ensuring no microscopic malignant cells are left behind in the resection tumor bed. Recent work from Haifler et al. has used Raman spectroscopy and AI to address this problem. A preliminary bench-top

model assessed six ex vivo samples of normal kidney parenchyma and six renal cell carcinoma sections and using a Bayesian machine learning classifier, sparse multinomial logistic regression, achieved an area under the ROC curve of 0.94 [13]. This data is limited by few data sets and internal validation; however, these analyses could be extrapolated to identify malignant tissue intraoperatively. The future prospect of using AI to render a 3D tumor reconstruction with intraoperative endoscopic vision and in situ pathology diagnostics may improve safety, efficacy, and accuracy of tumor resection during partial nephrectomy.

Urothelial Cancer

Cystoscopy is the gold standard in the diagnosis of bladder cancer. Ikeda et al. developed a convolutional neural network image classifier trained on 1.2 million general images, 8728 gastroscopic images, and finally 2102 cystoscopic images to detect bladder tumors. On a test set of 82/442 images containing bladder tumors, the algorithm performed with a 95.4% sensitivity and 97.6% specificity, which was comparable to the accuracy of an experienced urologist [14]. While AI has the potential to assist with the detection and possible automation of cystoscopy, other pursuits have applied AI to reduce the need for this invasive procedure. Sokolov et al. developed a random forest model based upon atomic force microscopy data of 5 cells from urine samples taken from 43 participants without bladder cancer and 25 bladder cancer patients. Using a 70/30 train-test split, they achieved 94% discriminatory accuracy [15]. Similar efforts by Sapre et al. used a support vector machine classifier trained on a microRNA panel from urine samples taken from patients without a history of bladder cancer, patients with a history of treated bladder cancer without evidence of recurrence, and patients with active bladder cancer to predict the presence of cancer cells. The classifier performed on an independent cohort with an AUC of 0.74 and sensitivity of 88%. The authors state their analysis could have reduced cystoscopy rates in the validation cohort by 30% [16]. While these methods remain experimental, AI systems have the potential to

increase the accuracy and reduce the cost and morbidity of office cystoscopy for the detection of bladder cancer.

Endourology

Endourology, the specialty of minimally invasive urological surgery, has been developed and fine-tuned over the last 40 years. One prominent division of endourology is the management of kidney stones. AI algorithms have been developed, principally in outcome prediction, for the management of nephrolithiasis.

One in 11 persons in the United States will experience kidney stones in their lifetime, and the initial presenting symptom is renal colic. In the acute setting, there is great clinical utility in predicting which patients will spontaneously pass their stones. This aids in patient guidance in whether surgical intervention is required. Solakhan et al. assessed 192 patients with ureteral stones presenting to an outpatient urology clinic. The authors created an artificial neural network model based on baseline stone size, body weight, pain score, ESR, and CRP. On a test set ($n = 30$), the model performed with 87.3% predictive accuracy of which patient's spontaneously passed their stones [17]. In a similar study, Dal Moro et al. assessed 402 patients presenting with renal colic using age, sex, body mass index, fever, previous urological treatments, previous expulsion of stones, duration of the symptoms (in hours), and dimension and position of the stone to build a support vector machine model to predict spontaneous stone passage within 6 months of the first presentation. Cross-validation showed the model successfully predicted spontaneous passage or surgical intervention with 84.5% sensitivity and 86.9% specificity [18]. In nephrolithiasis, predicting the need for surgical intervention is highly useful, as is prediction of surgical success in stone burden clearance pre-intervention.

Extracorporeal shock wave lithotripsy (ESWL) uses shock waves from an external source to fragment kidney stones. The primary measure of a successful ESWL is subsequent stone-free status. Well-validated traditional

statistical methods have shown body mass index, initial stone size, and skin to stone distance as key factors in predicting successful ESWL. Mannil et al. used a random forest classifier trained on preoperative data from 224 ESWL cases, including CT imaging-based stone 3D texture analysis, performed with an AUC of 0.81 when predicting successful ESWL [19]. However, a random forest analysis by Cui et al. of 459 ESWL cases with 19 clinically relevant inputs including stone 3D texture analysis performed with an AUC of 0.67 when predicting successful ESWL [20]. Further prediction in ability for ESWL to successfully clear stone burden will aid physicians and patients in choosing endoscopic vs. noninvasive treatment.

Percutaneous nephrolithotomy (PCNL) is the surgical intervention reserved for the largest kidney stones. The procedure involves accessing the kidney stone via a puncture through the back. Aminsharifi et al. developed an artificial neural network based on preoperative CT imaging and clinical data from 254 patients who underwent PCNL. The algorithm predicted stone clearance and perioperative blood transfusion with accuracies of 83% and 86%, respectively [21]. Recent work by Taguchi et al. uses AI to guide intraoperative percutaneous renal puncture in a pilot clinical trial. Using a table-mounted robotic system, the software determines the optimal trajectory for puncture using fluoroscopic images to guide the needle between the patient's skin and renal collecting system. After a ten-case learning curve, the author reported a puncture time of 2.8 min and no significant adverse events [22]. These data reflect the impact AI can have on improving patient surgical selection and its potential in intraoperative surgical planning and manipulation.

In addition to kidney stones, endourologists manage voiding dysfunction related to benign prostatic hypertrophy. Whangbo et al. developed a recurrent neural network to predict and measure time of voiding events from a smart wristband. The algorithm was trained on movement and tilt angle data collected from a three-axis accelerometer to recognize common three-step behavior for urination (forward movement, urination,

backward movement). The author has studied the accuracy of the algorithm compared to patient voiding diaries. They were able to predict voiding events in 30 participants over 60 days with 94.2% accuracy [23]. In summary, AI has the potential to impact nephrolithiasis in surgical planning and voiding dysfunction management.

Andrology

Andrology is a sub-specialty of urology that focuses on the medical and surgical management of all facets related to male reproductive health. This includes aspects of male reproductive urology such as infertility, reproductive endocrinology, and fertility preservation. It additionally includes sexual medicine components such as erectile dysfunction, Peyronie's disease, and priapism management. Andrology is strongly considered an area of medicine that artificial intelligence (AI) will greatly contribute toward the field's advancement. As the prevalence of infertility in the United States rises to approximately 7.3 million couples, and a male factor is identified in 50% of these patients, further tools are needed in the urologist's armamentarium. There are too many unknowns in this field of which time is of critical importance, and thus there has been foundation laid by many physicians and researchers for incorporating AI [24].

Prediction of Male Reproductive Potential

One of the most promising areas for AI in andrology is the prediction of male reproductive potential. While a reproductive urologist may be able to diagnose a patient after thorough male infertility workup, there have been efforts to identify patients earlier on in their reproductive years that may necessitate consultation. Additionally, it may inform practitioners to focus on a male factor without the need for expensive diagnostics. An early emphasis has been on epigenetic research to uncover possible environmental or lifestyle factors that affect a patient's semen parameters, through the use of AI. Girela et al. preprocessed human variable data derived from a questionnaire obtained

from 100 healthy volunteers, such as socio-demographic data, cigarette and alcohol use, body mass index, and general health status with a decision tree. Using these data, the team developed a multilayer perceptron (MLP) artificial neural network (ANN) that predicted sperm concentration with 90% accuracy (sensitivity, 95.45%, and specificity, 50%) and motility with 82% accuracy (sensitivity, 89.29%, and specificity, 43.75%) through these factors [25]. Incorporation of these AI networks into annual practitioner visits may lead to more timely fertility interventions.

Identification of seminal biomarkers that may implicate fertility potential has also been researched. In particular, seminal zinc and leptin levels have been correlated as possible surrogates for male infertility, as an appropriate threshold may be required for basic sperm function. Ma et al. identified seminal leptin as a potential biomarker to be utilized alongside an artificial neural network (ANN) model in determining sperm retrieval success in patients with non-obstructive azoospermia (NOA). Two hundred eighty patients with NOA were trained with various inputs such as testicular volume, semen volume, hormone levels, seminal leptin levels, and semen parameters. The ANN model had a AUC = 0.83, which was better than single independent variables [26]. Vickram et al. validated a backpropagation neural network, derived from a training set of 177 semen samples, that was able to predict seminal biochemical parameters including protein (mean absolute error = 0.025), fructose (mean absolute error = -0.080), glucosidase (mean absolute error = 0.166), and zinc levels (mean absolute error = -0.057) utilizing just semen parameters [27]. While semen parameters and hormonal levels comprise the diagnostic foundation of the male infertility workup, these studies utilizing AI show promise to enable quantification of reproductive potential from seminal biomarkers.

In patients who have been identified to be azoospermic, the next step of the male infertility workup is undergoing chromosomal testing to determine possible genetic etiologies. These tests prove to be expensive and require specialized labs that cause a time lag in the infertility cycle. Akinsal et al. trained an ANN with 310 azoospermic patient

data, physical exam, hormonal levels, and ejaculate volume to identify patients with 95% accuracy who would most likely need further investigation [28].

The mark of achieving full male fertility potential begins with penetration of sperm into the ova. This is considered conception success, and Niederberger et al. attempted to predict this fertility potential just by semen analysis (SA) data. They demonstrated that neural networks showed promising results in predicting sperm penetration success rates in bovine cervical mucus on various ova based just on the SA. The neural network was able to correctly classify 67.8% of zona-free hamster egg penetration assays (1416 assays in training set) and 80% of bovine cervical mucus sperm penetration results (139 assays in training set) [29]. With the advent of artificial reproductive technologies (ART) such as intracytoplasmic sperm injection (ICSI), sperm selection among a heterogenous specimen sample has risen to be of paramount importance as penetration of the ova is bypassed. Intrinsic qualities that are currently unable to be fully evaluated by the human eye may be of integral importance for sustained conception success and resultant live birth. Current modalities include a broad set of laboratory bench manipulation such as microfluidic chamber challenge, protein binding, and morphological examination. Preliminary research as presented by Takeshi et al. at the 2018 American Society for Reproductive Medicine Conference demonstrated an ability to differentiate 8010 spermatozoa and 25,522 non-spermatozoa cells through AI. Mirsky et al. isolated sperm cell heads of 1400 human sperm cells through interferometric phase microscopy and utilized this data to train a support vector machine to automatically classify sperm as good and bad morphology for possible use in in vitro fertilization. The SVM had a ROC of 88.59% and AUC of 88.67%, with precisions at 90% or higher [30]. Further research utilizing AI in sperm selection is needed, especially as progress and promise have been shown from the embryo selection side [31].

Semen Analyses

Artificial intelligence has had an impact on the realm of diagnostics in reproductive urology. The mainstay diagnostic of an infertility workup

is the semen analysis, a test that is traditionally performed in an andrology laboratory with manual counting by a technician or computer-assisted sperm analysis (CASA) system. This has been a labor-intensive and time-consuming diagnostic with potential subjective operator bias. Agarwal et al. have demonstrated automation of SA by utilizing a novel optical microscope that is built upon a proprietary artificial intelligence algorithm. By comparing the results of a manual and AI count on 135 semen samples, they observed high degree of correlation ($r > 0.9$) in sperm concentration, motility, and pH results [32]. Advances in consumer technology have led to an increase in home semen analysis kits over the past few years. In particular, digital home kits that utilize a smartphone camera and device adaptor can provide consumers with semen parameter results within minutes. While initial versions were limited in only providing a few parameters, the supplementation of AI computer vision technology has expanded accuracy in measuring both sperm motility and morphology. Tsai et al. evaluated this AI system by comparing automatically generated semen parameters to objective grading of sperm quality by experts and found automatically generated concentration of total sperm ($r = 0.65$, $p < 0.001$), concentration of motile sperm ($r = 0.84$, $p < 0.001$), and motility percentage ($r = 0.90$, $P < 0.001$) correlated well with expert grades [33].

Predicting Sperm Retrieval Success Rates

In infertile male patients found to be with non-obstructive azoospermia, sperm retrieval is the procedure that must be undergone to possibly find sperm for ART. The success of a microdissection testicular sperm extraction (micro-TESE) has been reported to be approximately 63%. Ramasamy et al. formulated an ANN that predicted sperm retrieval success rates by training the system with preoperative clinical data and operative results from 1026 men who underwent micro-TESE. The ANN was able to accurately predict 59.4% of the operative outcomes (AUC 0.64). This algorithm with moderate discrimination ability may still be clinically useful, especially when counseling patients who are

unwilling or unable to have a micro-TESE. These analyses demonstrate promise for AI in aiding patients and physicians in managing expectations of surgical success [34].

Predicting Surgical Shunt Intervention for Priapism Management

Priapisms are defined as prolonged erections, usually greater than four, and may have various etiologies. Classification is divided into either ischemic, non-ischemic, or recurrent. Emergent management of ischemic priapism is important as it, if left untreated, may possibly lead to possible male impotence. Identification of those that may require escalation to surgical shunting would save valuable time trialing less invasive treatments that may not resolve the priapism. Masterson et al. performed a retrospective study of 364 priapism encounters at a county hospital, collecting data on patient age, race/ethnicity, type of priapism, etiology, duration of erection prior to presentation, length of hospital encounter, prior episode of priapism, interventions attempted, complications, escalation to operative intervention, urological follow-up, and patient risk factors. Three hundred thirty-four cases were determined to be ischemic etiologies. The team utilized a random forest machine learning algorithm to train the data set and were able to create an online risk calculator to accurately identify those who require surgical shunting or not 87.2% of the time with AUC 0.76. Positive predictive value remained low at 44.5%. It was identified that longer duration of priapism was a notable factor for surgical shunting. Further studies are required to increase the validity of the study, but its initial results show promise for AI in predicting surgical intervention for ischemic priapism [35].

Conclusion

Artificial intelligence is poised to make a significant impact on the field of urology, allowing for further improvements in patient care. The possibilities of AI applications are only scratching the surface, as exponentially more literature is published year after year. The broad scope of AI urologic application includes

oncology, endourology, and andrology. With the urology track record in innovation, AI will see many of its potential reflected in urologic practice over the next decade.

References

1. Scardino PT. Urology: a long history of innovation. *Nat Clin Pract Urol.* 2008;5(2):59.
2. Partin AW, Yoo J, Carter HB, Pearson JD, Chan DW, Epstein JI, et al. The use of prostate specific antigen, clinical stage and Gleason score to predict pathological stage in men with localized prostate cancer. *J Urol.* 1993;150(1):110–4.
3. Yates DR, Vaessen C, Roupret M. From Leonardo to da Vinci: the history of robot-assisted surgery in urology. *BJU Int.* 2011;108(11):1708–13; discussion 1714.
4. Auffenberg GB, Ghani KR, Ramani S, Usoro E, Denton B, Rogers C, et al. askMUSIC: leveraging a clinical registry to develop a new machine learning model to inform patients of prostate cancer treatments chosen by similar men. *Eur Urol.* 2019;75(6):901–7.
5. Nagpal K, Foote D, Tan F, Liu Y, Chen P-HC, Steiner DF, et al. Development and validation of a deep learning algorithm for Gleason grading of prostate cancer from biopsy specimens. *JAMA Oncol.* 2020;6(9):1372–80.
6. Steiner DF, Nagpal K, Sayres R, Foote DJ, Wedin BD, Pearce A, et al. Evaluation of the use of combined artificial intelligence and pathologist assessment to review and grade prostate biopsies. *JAMA Netw Open.* 2020;3(11):e2023267.
7. Schelb P, Kohl S, Radtke JP, Wiesenfarth M, Kickingereder P, Bickelhaupt S, et al. Classification of Cancer at prostate MRI: deep learning versus clinical PI-RADS assessment. *Radiology.* 2019;293(3):607–17.
8. Fehr D, Veeraraghavan H, Wibmer A, Gondo T, Matsumoto K, Vargas HA, et al. Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images. *Proc Natl Acad Sci U S A.* 2015;112(46):E6265–73.
9. Hung AJ, Chen J, Che Z, Nilanon T, Jarc A, Titus M, et al. Utilizing machine learning and automated performance metrics to evaluate robot-assisted radical prostatectomy performance and predict outcomes. *J Endourol.* 2018;32(5):438–44.
10. Hung AJ, Chen J, Ghodousipour S, Oh PJ, Liu Z, Nguyen J, et al. A deep-learning model using automated performance metrics and clinical features to predict urinary continence recovery after robot-assisted radical prostatectomy. *BJU Int.* 2019;124(3):487–95.
11. Heller N, Isensee F, Maier-Hein KH, Hou X, Xie C, Li F, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: results of the KiTS19 challenge. *Med Image Anal.* 2021;67:101821.

12. Hattab G, Arnold M, Strenger L, Allan M, Arsentjeva D, Gold O, et al. Kidney edge detection in laparoscopic image data for computer-assisted surgery: kidney edge detection. *Int J Comput Assist Radiol Surg.* 2020;15(3):379–87.
13. Haifler M, Pence I, Sun Y, Kutikov A, Uzzo RG, Mahadevan-Jansen A, et al. Discrimination of malignant and normal kidney tissue with short wave infrared dispersive Raman spectroscopy. *J Biophotonics.* 2018;11(6):e201700188.
14. Ikeda A, Nosato H, Kochi Y, Negoro H, Kojima T, Sakanashi H, et al. Cystoscopic imaging for bladder cancer detection based on stepwise organic transfer learning with a pretrained convolutional neural network. *J Endourol.* 2020. <https://doi.org/10.1089/end.2020.0919>. Online ahead of print.
15. Sokolov I, Dokukin ME, Kalaparthi V, Miljkovic M, Wang A, Seigne JD, et al. Noninvasive diagnostic imaging using machine-learning analysis of nano-resolution images of cell surfaces: detection of bladder cancer. *Proc Natl Acad Sci U S A.* 2018;115(51):12920–5.
16. Sapre N, Macintyre G, Clarkson M, Naeem H, Cmero M, Kowalczyk A, et al. A urinary microRNA signature can predict the presence of bladder urothelial carcinoma in patients undergoing surveillance. *Br J Cancer.* 2016;114(4):454–62.
17. Solakhan M, Seckiner SU, Seckiner I. A neural network-based algorithm for predicting the spontaneous passage of ureteral stones. *Urolithiasis.* 2020;48(6):527–32.
18. Dal Moro F, Abate A, Lanckriet GRG, Arandjelovic G, Gasparella P, Bassi P, et al. A novel approach for accurate prediction of spontaneous passage of ureteral stones: support vector machines. *Kidney Int.* 2006;69(1):157–60.
19. Mannil M, von Spiczak J, Hermanns T, Poyet C, Alkadhi H, Fankhauser CD. Three-dimensional texture analysis with machine learning provides incremental predictive information for successful shock wave lithotripsy in patients with kidney stones. *J Urol.* 2018;200(4):829–36.
20. Cui HW, Silva MD, Mills AW, North BV, Turney BW. Predicting shockwave lithotripsy outcome for urolithiasis using clinical and stone computed tomography texture analysis variables. *Sci Rep.* 2019;9(1):14674.
21. Aminsharifi A, Irani D, Pooyesh S, Parvin H, Dehghani S, Yousofi K, et al. Artificial neural network system to predict the postoperative outcome of percutaneous nephrolithotomy. *J Endourol.* 2017;31(5):461–7.
22. Taguchi K, Hamamoto S, Kato T, Iwatsuki S, Etani T, Okada A, et al. Robot-assisted fluoroscopy-guided renal puncture for endoscopic combined intrarenal surgery: a pilot single-centre clinical trial. *BJU Int.* 2020;127:307.
23. Whangbo T-K, Eun S-J, Jung E-Y, Park DK, Kim SJ, Kim CH, et al. Personalized urination activity recognition based on a recurrent neural network using smart band. *Int Neurourol J.* 2018;22(Suppl 2):S91–100.
24. Chu KY, Nassau DE, Arora H, Lokeshwar SD, Madhusoodanan V, Ramasamy R. Artificial intelligence in reproductive urology. *Curr Urol Rep.* 2019;20(9):52.
25. Girela JL, Gil D, Johnsson M, Gomez-Torres MJ, De Juan J. Semen parameters can be predicted from environmental factors and lifestyle using artificial intelligence methods. *Biol Reprod.* 2013;88(4):99.
26. Ma Y, Chen B, Wang H, Hu K, Huang Y. Prediction of sperm retrieval in men with non-obstructive azoospermia using artificial neural networks: leptin is a good assistant diagnostic marker. *Hum Reprod Oxf Engl.* 2011;26(2):294–8.
27. Vickram AS, Kamini AR, Das R, Pathy MR, Parameswari R, Archana K, et al. Validation of artificial neural network models for predicting biochemical markers associated with male infertility. *Syst Biol Reprod Med.* 2016;62(4):258–65.
28. Akinsal EC, Haznedar B, Baydilli N, Kalinli A, Ozturk A, Ekmekcioğlu O. Artificial neural network for the prediction of chromosomal abnormalities in azoospermic males. *Urol J.* 2018;15(3):122–5.
29. Niederberger CS, Lipshultz LI, Lamb DJ. A neural network to analyze fertility data. *Fertil Steril.* 1993;60(2):324–30.
30. Mirsky SK, Barnea I, Levi M, Greenspan H, Shaked NT. Automated analysis of individual sperm cells using stain-free interferometric phase microscopy and machine learning. *Cytom Part J Int Soc Anal Cytol.* 2017;91(9):893–900.
31. Curchoe CL, Bormann CL. Artificial intelligence and machine learning for human reproduction and embryology presented at ASRM and ESHRE 2018. *J Assist Reprod Genet.* 2019;36(4):591–600.
32. Agarwal A, Henkel R, Huang C-C, Lee M-S. Automation of human semen analysis using a novel artificial intelligence optical microscopic technology. *Andrologia.* 2019;51(11):e13440.
33. Tsai VF, Zhuang B, Pong Y-H, Hsieh J-T, Chang H-C. Web- and artificial intelligence-based image recognition for sperm motility analysis: verification study. *JMIR Med Inform.* 2020;8(11):e20031.
34. Ramasamy R, Padilla WO, Osterberg EC, Srivastava A, Reifsnyder JE, Niederberger C, et al. A comparison of models for predicting sperm retrieval before microdissection testicular sperm extraction in men with nonobstructive azoospermia. *J Urol.* 2013;189(2):638–42.
35. Masterson TA, Parmar M, Tradewell MB, Nackeeraan S, Rainer Q, Blachman-Braun R, et al. Using artificial intelligence to predict surgical shunts in men with ischemic priapism. *J Urol.* 2020;204(5):1033–8.



Artificial Intelligence in Trauma and Orthopedics

63

Roshana Mehdian and Matthew Howard

Contents

Introduction	874
Diagnostics	875
Musculoskeletal Image Scheduling and Protocoling	876
Musculoskeletal Image Acquisition	876
Musculoskeletal Image Interpretation	876
Intraoperative and Robotics	878
Semiautonomous Intraoperative Robotics	879
Autonomous Intraoperative Robots	880
Continued Adoption of Robotics	881
Predictive Analytics	881
Orthopedic Databases	881
Predicting Disease Onset and Degree	882
Postoperative Complications and Rehabilitation	882
Conclusion	883
References	883

Abstract

This chapter will explore artificial intelligence (AI) in trauma and orthopedics (orthopedics). Orthopedics is a branch of surgery that focuses on the prevention of musculoskeletal pathology and the correction and restoration of form

and function of these structures. Orthopedics is fertile ground for adoption of technological innovations, including artificial intelligence, where small gains in the treatment of one condition can lead to improved outcomes for some of the largest patient populations in medicine. Orthopedics is well suited to innovation and the application of AI as it has clear pathways for common diseases and is a highly technical field with constant technical innovation. This chapter will review several of the key applications of AI in orthopedics including diagnostics, intraoperative robotics, and predictive analytics.

R. Mehdian (✉)
St Georges Hospital London NHS, London, UK

M. Howard
Musgrove Park Hospital, Taunton, UK
e-mail: Matthew.howard@nhs.net

Keywords

Trauma · Orthopedics · Fracture · Arthroplasty · Machine learning · Deep learning · Artificial intelligence · Robotics · Surgery

Introduction

This chapter will explore artificial intelligence (AI) in trauma and orthopedics (orthopedics).

Orthopedics is a branch of surgery concerned with the musculoskeletal system, including the extremities, spine, and their associated structures. The specialty focuses on the prevention of musculoskeletal pathology and the correction and restoration of form and function of these structures. Trauma, in the context of orthopedics, refers to a wide spectrum of acute injuries of the musculoskeletal system caused by an external force, for instance, fractures caused by a car accident or tendon ruptures caused by landing a jump awkwardly. Orthopedics also consists of elective care which is the prevention, diagnosis, and management of chronic musculoskeletal conditions such as arthritis.

As well as impacting health outcomes, including a patient's quality of life, musculoskeletal conditions are a significant burden on healthcare systems. In the UK 25% of all surgical interventions are musculoskeletal and account for 4.7 billion of NHS spending each year [1]. By impacting individual productivity, they also have an indirect impact on a country's economy. In the USA one in two adults is thought to be affected by musculoskeletal problems costing an estimated \$213 billion in treatment and economic productivity, approximately 1.4% of US GDP [2]. Consequently, the need for delivery of reproducible and effective orthopedic care means it is fertile ground for adoption of technological innovations, including artificial intelligence; small gains in the treatment of one condition can lead to improved outcomes for some of the largest patient populations in medicine and significant economic savings.

Beyond issues of scale and cost, multiple features of orthopedics mean that it is particularly

well suited to innovation and the adoption of artificial intelligence:

1. Orthopedics is a highly technical field, characterized by consistent technological development. Orthopedic practice consists of the daily use of technology ranging from prosthetic implants to robotics. Its practitioners are required to develop sound technical understanding to properly assess, plan, and deploy the use of these technologies in the treatment of their patients. The orthopedic surgeon workforce is therefore poised for relatively rapid adoption of new technologies once they are proven effective.
2. Orthopedics is used to close working relationships with technology-creating industries. These close ties with industry mean that there are established mechanisms for real-time feedback and clinician-industrial partnership development.
3. Orthopedic conditions have well described diagnostic and treatment pathways, which lend themselves to optimization on a large scale.
4. Many common orthopedic operations are reproducible and are associated with highly effective outcomes, for instance, hip replacements [3]. The mature and established status of such procedures confers a relative uniformity for the research and development of new technologies and advances in optimization.
5. The orthopedic community is pioneer of "big data." Orthopedics was the first medical specialty to institute national and international databases that hold large repositories of procedure-specific data [4]. A key roadblock to the development of artificial intelligence can be access to sufficient data to optimize accuracy and performance. The maturity of extant orthopedic databases can offer fertile datasets for the advancement of AI applications.

Due to these factors, the research and development of a range of AI-based techniques and applications in orthopedics has been advancing rapidly. A recent review has mapped the breadth of these applications across orthopedic practice and the AI techniques employed (Fig. 1) [5].



Fig. 1 Bubble chart diagram showing applications of machine learning (ML) in orthopedics by ML technique. (Published in a recent review from: Machine

Learning in Orthopedics – A Literature Review, Cabitza et al. Copyright. Reproduced with permission by Creative Commons Attribution 4.0 international [5])

This chapter will explore the use of artificial intelligence in some key domains in orthopedics, diagnostics, robotics, and predictive analytics including orthopedic databases and postoperative complications.

Diagnostics

Orthopedics diagnosis is heavily reliant on radiological imaging, with the most common tests used to diagnose and quantify orthopedic images being radiographs (X-ray), ultrasound, computed

tomography (CT), and magnetic resonance imaging (MRI). A diagnostic image is most commonly the key instigator for orthopedic referral and treatment. The great majority of injuries are first seen by a primary care practitioner or the emergency department, who orders the initial imaging. In these cases it is not uncommon for insufficient radiological views to have been ordered to allow for diagnosis and quantification or ruling out a specific injury. An orthopedic specialist may then make another order for additional views or a more sensitive scan at a later time point in the patient journey. This can lead to a delay in starting

treatment or missed fractures, resulting in poorer outcomes for patients, increased need for complex delayed intervention, and increased cost [6, 7]. Missed or occult fractures also account for a significant proportion of medicolegal claims [8].

Examples of commonly missed fractures with associated morbidity include scaphoid fractures, leading to increased incidence of avascular necrosis [6]; up to 20% of these are missed radiologically, and capture of these fractures can be improved with appropriate views, serial imaging, or a protocol to urgent CT or MRI scan [6, 9]. Lisfranc fracture dislocations in the foot are missed in 20% of cases [7]. If the wrong X-ray views are taken, these can be missed in up to 50% of cases [10]. In both of these examples, early recognition of the mechanisms of injury that lead to the injury, the correct image protocol, and a low threshold to use imaging with higher diagnostic capability can reduce missed fractures and therefore morbidity and cost significantly.

The applications of AI in orthopedic diagnostics can help reduce the likelihood of missed injuries and ensure timely treatment. Applications range from improving upstream functions such as faster image acquisition and improved protocoling to downstream applications such as automated image analysis and interpretation.

Musculoskeletal Image Scheduling and Protocoling

Improved image scheduling and protocoling in orthopedics can lead to faster diagnosis. Natural language processing (NLP), a form of machine learning (ML) that enables computers to derive meaning from unstructured human input, has shown early promise in this area.

The technology may enable processing and triage of electronic health records and imaging requests to schedule and protocol accordingly, ensuring appropriate imaging and views. Deep learning convolutional neural networks (CNNs) have already been developed to choose between tumor or routine musculoskeletal MRI protocols based on clinical information and indication with an accuracy level of 94% [11]. One group

developed a natural language classifier that was able to extract the need for contrast for musculoskeletal MRIs with an accuracy of 83% [12]. Applications of this technology have also been able to extract radiologist recommendations for follow-up imaging from reports [13]. Examples like this could act to streamline the imaging pathway and reduce time between scans for many orthopedic patients.

Musculoskeletal Image Acquisition

AI has been applied to expedite the acquisition of knee MRI. In acute soft tissue knee injury, it is important for patients to receive treatment in a timely manner so as not to suffer poorer outcomes [14–16]. Patients often present to ED, where radiographs are taken and a referral made to orthopedic clinic. Complete assessment of suspected ligamentous or meniscal injury of the knee necessitates an MRI, which is often ordered by the orthopedic specialist once the patient is seen in clinic, as these are costly and require considerable time to scan. Availability of faster acquisition MRI could reduce costs and scan time and therefore could be ordered earlier in the patient journey, increasing the likelihood of timely treatment and better outcomes. Techniques for accelerated acquisition have been available since the 2000s, but reconstructed images tend to contain artifacts – leading to a trade-off between quality and speed. Using ML to help with reconstruction of accelerated sequences, such as knee MRI, can reduce artifacts and therefore improve the accuracy of these scans. Studies have shown promising results in preserving the quality of MRI and have outperformed standard reconstruction methods [17].

Musculoskeletal Image Interpretation

AI in radiological interpretation is a popular field and has been applied to musculoskeletal imaging using convolutional neural networks (CNNs), a form of machine learning. CNNs are well suited to carry out analytical tasks such as detection (fracture detection, soft tissue knee pathology

detection or spinal pathology), classification (osteoarthritis classification, spinal deformities), or segmentation (cartilage or meniscus) [18].

Fracture Detection

Multiple studies have been published demonstrating the application of deep learning (DL) algorithms to radiographs for automated fracture detection. Systems designed to look at a variety of peripheral radiographs have shown accuracies of around 83% [19]. A DL algorithm trained specifically to identify distal radius fractures achieved even more impressive results with a sensitivity of 90% and specificity of 88% [20].

A recent review of the musculoskeletal fracture detection systems available found several AI tools that have been able to identify orthopedic injuries on X-ray with greater accuracy than physicians who are not orthopedic surgeons or radiologists [18]. A few studies have shown accuracy approaching or equal to that of orthopedic surgeons and radiologists, but the greatest results are achieved when used as clinical decision support to assist in these specialists' interpretations rather than used alone [18, 20]. The work of orthopedic surgeons or radiologists involves more than fracture detection; they also consider the patients' clinical state, compare historical images, and communicate with doctors of other specialisms; for this reason, fracture detection tool developers and healthcare providers have been unwilling to shift this workload and therefore liability from the clinician to AI tools. Instead, these computer-assisted detection (CAD) tools can be used to augment the interpretation of the non-specialist (non-orthopedic or radiology specialist) or busy clinician.

Knee Pathology Detection and Segmentation

Several pathology detection tools have been developed for the identification of soft tissue knee injuries on MRI. These have focused primarily on anterior cruciate ligament (ACL) tears or meniscal tears. ACL tears are a common orthopedic injury that require timely treatment, often surgical, to prevent undesirable sequelae such as meniscal tears or arthritic wear. Meniscal tears,

similarly, are common orthopedic injuries which often require surgery. Both of these pathologies generally require MRI to confirm a diagnosis and proceed to surgery. Given their prevalence, these injuries have been the focus of a number of AI detection research projects [21–23].

A CNN developed to identify and classify ACL tears as no tear, partial, or complete had a similar specificity but lower sensitivity than specialist musculoskeletal radiologists. When the CNN was used as a CAD tool by general radiologists, it increased specificity by 4.8% when compared with radiologist-only performance [21, 24]. CNNs developed to look only at no tear versus complete tear showed no statistical difference when compared to radiologists [23, 24]. CNNs developed for the detection and classification of meniscal tears have yielded good but slightly inferior results and have not yet been able to emulate the sensitivity and specificity of expert radiologists [21, 24]. The interpretation of meniscal tears also requires classification of type of tear as each type can be treated differently, which CNNs have not yet been able to achieve. While there are promising results, the deficiency in classification ability and detection of other concomitant soft tissue knee injury, which is often present, means that these tools are still some way off mainstream adoption.

Osteoarthritis Detection and Cartilage Segmentation

Osteoarthritis (OA) is commonly identified and confirmed after clinical assessment with an X-ray. Historically, radiographs have been the imaging of choice, but with the advent of cartilage preservation treatments and unicompartmental arthroplasty, more detailed imaging, such as MRI, is increasingly being used to quantify the degree and specific locations of most wear. AI has been developed for OA detection in both these imaging modalities.

CNNs to detect and classify cartilaginous defects and osteoarthritis based on the Kellgren and Lawrence system (a method of classifying osteoarthritis using five grades) in knee radiographs have achieved multiclass classification accuracy of up to 66.7% [25]. One model that used a combined approach of a CNN and

demographic data was able to achieve knee model sensitivities for no, mild, moderate, and severe OA of 83.7%, 70.2%, 68.9%, and 86.0%, with corresponding specificities of 86.1%, 83.8%, 97.1%, and 99.1% [24–26]. A model designed to detect the presence of hip OA from radiographs exhibited similar capabilities to that of experienced radiologists [25]. An MRI-based CNN assessing knee OA was found to show improved sensitivity but lower specificity when compared with radiologists [27]. There is some way to go to attain consistent results comparable to the expert radiologist or orthopedic surgeons, but the ongoing development of OA CAD technology could one day help in alleviating the burden of this highly prevalent disease.

Orthopedic Implant Detection

Some of the most challenging and costly procedures in orthopedics are revision arthroplasties. To appropriately deal with a problem arthroplasty, the orthopedic surgeon must first identify which implant has been used in the primary procedure. As operations are often performed at other distinct centers, with implant data often inaccessible on disparate electronic systems, it can be difficult to correctly identify implants. One survey found 88% of surgeons reported this to be a critical barrier in revision arthroplasty surgical planning and execution [28]. AI tools have been employed to classify implants thereby potentially reducing risk of delay and surgeon hours used in identifying the implant and likelihood of inappropriate surgical equipment in the operating theater. A convolutional neural network deep learning algorithm developed across four sites in the USA is able to use plain radiographs to detect and classify hip arthroplasty implants. The accuracy of the model on its test set was an accuracy of 99.6% [29]. One model trained to identify radiographs of three commonly used hip implants achieved an accuracy of 100% [30]. Such identification tools are now commercially available and can identify arthroplasties from multiple joints including knee, shoulder, and hip – though accuracy of these systems is unclear due to a lack of published research.

Orthopedic Oncology Detection

Bone tumors are relatively rare and therefore present a challenge for both radiologists and orthopedic surgeons. Work on CAD for bone tumors began as early as the 1960s with the work of Lodwick, a musculoskeletal radiologist. Lodwick developed a computer program based upon Bayes formula that returned probability of a bone tumor diagnosis using demographic data and radiographic features. It predicted correctly 77.9% of the time among eight bone tumor types [31]. More recent work has improved further on these results with the use of ML CNNs to develop CAD tools. Tools have been developed to characterize bone lesions as malignant versus benign based on radiographs [32]; classify bone destruction, staging, and grading of bone lesions using a decision tree tool on radiographic images [33]; and classify vertebral body bone lesions on cross-sectional CT imaging benign or malignant using a random forest classification [34].

Intraoperative and Robotics

Intraoperative applications of artificial intelligence in orthopedics mainly come in the form of computer-assisted navigation, computer-assisted surgery (CAN/CAS), or robotic technology. Although available in some guise since the late 1980s, this field has been relatively slow to take off and has only recently reached mainstream practice [35]. CAN systems use referencing methods to reconstruct virtually the desired operative area usually with the aid of preoperative CT, surgeon assist planning, or fluoroscopy images; they then provide the surgeon with real-time intraoperative feedback [36]. Robotic systems generally consist of computer-guided mechanical instruments that in theory allow for more accurate and reproducible placement of tools and implants.

These systems can also be divided into passive (CAN), semi-active (robot-assisted), and active (autonomous robotic) [35], with only the latter two truly falling under the scope of artificial intelligence.

Robotic autonomy can be thought of as a spectrum, with the most rudimentary robots lacking

autonomy and pre-programmed to run a set of repetitive limited functional tasks. Fully autonomous robots are able to perform complex tasks without human intervention. Between these two extremes are semiautonomous robots which may incorporate artificial intelligence algorithms into specific elements of their function or may use AI to continuously improve their performance with adaptive learning. The degree of autonomy is defined by the interplay of three factors: mission complexity, human independence, and environmental difficulty as described by ALFUS framework [37].

Fields within orthopedics that have been at the forefront of taking up the development and use of robotics have predominantly involved alignment and positioning, such as the implantation of arthroplasty components, placement of spinal pedicle screws, and bone or soft tissue resection margins in oncology surgery [38, 39]. The focus of this section will be on arthroplasty robotic systems, as an example use case.

Passive, or CAN, systems usually have multiple sensors which can track the movements of surgical instruments and assess bone morphology and alignment. They provide detailed information to the surgeon, acting as a real-time referencing system [40] (Fig. 2).

Semiautonomous Intraoperative Robotics

The ability for a robot to perceive its environment – “perception” – requires AI algorithms. A robot that has perceived its environment in order to locate and carry out a task with an adaptive response uses AI for perception and output directions, though the mechanization to carry out the task does not require AI. If the resulting output is used to guide the surgeon to perform the task, rather than perform it itself, the robot is semiautonomous. Semi-active or robot-assisted systems can act to instruct and constrain actions that the surgeon is performing, for example, during ligament balancing or ensuring that bone will only be resected within a pre-defined cutting zone in a unicompartmental knee replacement [42].

In orthopedics, knee arthroplasty is a popular area for the development of robotics; these robots are used to guide a bone cut along an optimal pre-prescribed path decided from preoperative CT by the surgeon with assistance from the planning software. The robot will then perceive its environment and constrain the movements of the surgeon using passive assistance (pushing the surgeons and saw away from a path) or active assistance (guiding the surgeon toward a path). They can also make real-time assessments of soft tissue balance and feed this back to the surgeon [35, 43, 44].

As well as increased accuracy, these systems have also been shown to have the potential to reduce the need for fluoroscopy during procedures. In some surgical specialties, though not yet in orthopedics, they can facilitate remote-operating – with the surgeon working in a different location to the patient [45]. Although these machines have been shown to improve the alignment and positioning of components in arthroplasty procedures [46], there is conflicting evidence to show whether this converts to improved clinical outcome or functional scores at short- or long-term follow-up [43, 47, 48], meaning some have felt their use (and significant cost) has been difficult to justify in widespread practice [49]. Few knee surgeons would refute however that the key to short- and long-term success in knee arthroplasty is achieving optimal alignment with good ligament balancing, and while subjective feel comes with experience, there can be no negative to objective real-time data. This can be achieved during the procedure with a combination of robotic-assisted cutting and computer-assisted ligament balancing. A recent study using the Australian Orthopaedic Association National Joint Replacement Registry demonstrated that at 3 years, a certain brand of robot-assisted unicompartmental knee arthroplasty had significantly lower overall revision rates compared to other types of non-robotically assisted procedures at 3 years. The revision for aseptic loosening was lower, supporting the theory that robotic assistance improves alignment; however, revision for infection was significantly higher in the robotic-assisted group, which could be due to longer surgical time [43, 50].

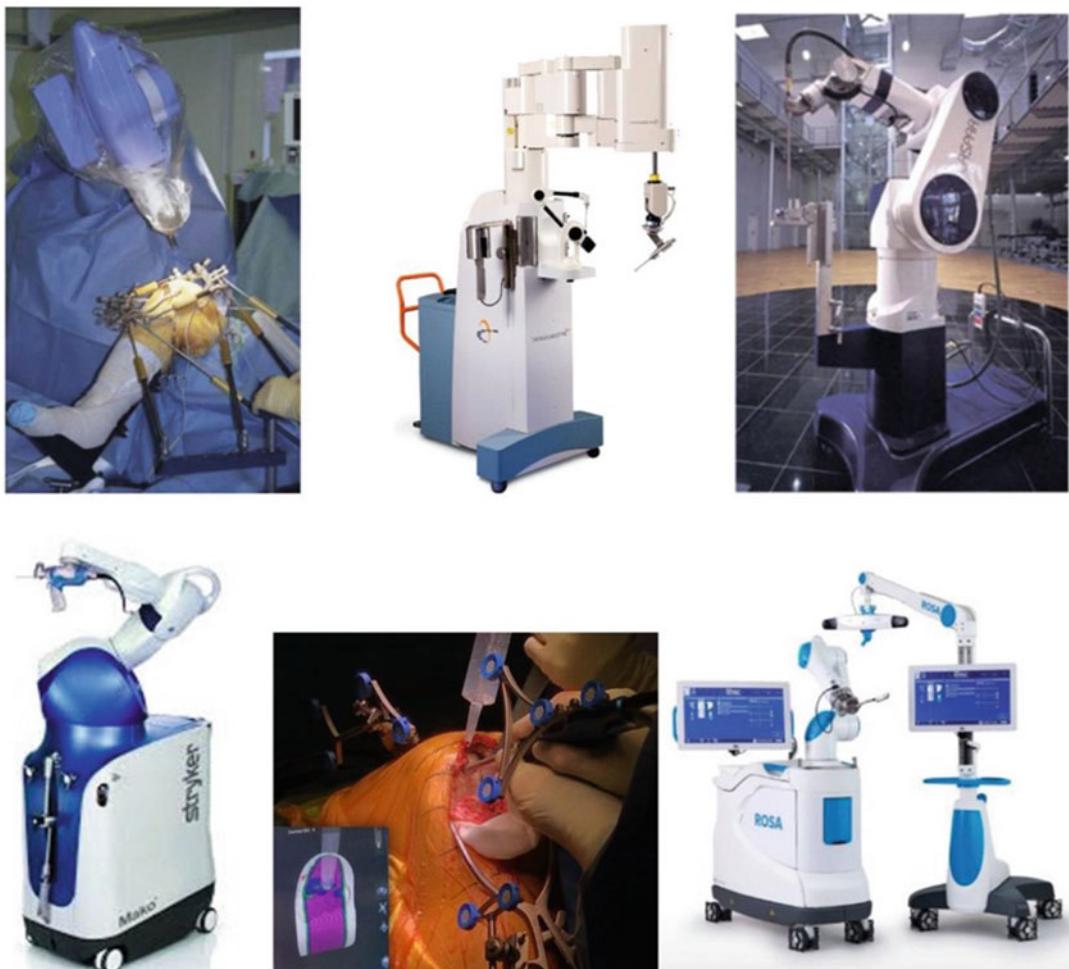


Fig. 2 Autonomous and semiautonomous robotic systems: top 3 images are autonomous robots (Robodoc, Caspar). The bottom three robots are semiautonomous (Mako, Navio, Rosa). (From: New Technologies in Knee

Arthroplasty: Current Concepts, by Batailler, C [41]. Image reproduced with permission under the Creative Commons Attribution 4.0 International license)

Autonomous Intraoperative Robots

Active, or autonomous, systems are what the public often perceive when they hear the phrase “robotic surgery.” The reality is that a robotic system performing even a part of a procedure without a guiding surgeon’s hand is still some way off from being accepted in clinical practice [36], both from a technical and ethical perspective. Although patient concern might be considered a potential barrier to uptake for robotic surgery in general, recent research has shown that almost half of patients would not mind at

least part of their surgery being performed by autonomous robotic technology [51].

The only long-term follow-up trial for robotic knee surgery is for an autonomous system. The study required a 10-year minimum follow-up of both conventional and robot-assisted total knee replacement (TKA). This found that there were no differences between the autonomous robot-assisted TKA and non-robotic TKA when comparing functional outcome scores, aseptic loosening, survivorship, and complications. However, the authors could not recommend its use citing the lack of benefit over conventional means and increased outlays [52].

Continued Adoption of Robotics

Globally, the number of knees implanted using robotic technology are currently very small compared to total conventional operations, but data will continue to emerge, particularly as global national joint registries have now started to collect this data.

Although huge technical advances have been made, it is difficult to see these technologies becoming widespread until clear long-term functional and clinical benefit can be proven, to outweigh the significant costs involved – initial outlay for a single robotic system can be in the millions of pounds, with annual maintenance running into hundreds of thousands [13]. It would seem that for these costs to significantly reduce, large volume of uptake would be needed to create some economies of scale. Reaching this uptake will be difficult due to the slow-growing cohort of patients who have been operated on using this technology that can be used as a study population, with more time needed to fully assess any differences in long-term outcome. Currently there is a paucity of data on the long-term outcomes of the techniques that are currently available, but there is however a reasonably steady stream of positive literature in short-term studies [43, 53, 54].

A further factor not to be overlooked is that there is clearly a learning curve involved for the surgeon using this technology [50], and it is important that appropriate infrastructure is developed to allow adequate training of current and future surgeons to be able to embrace these systems safely [55]. Equally, there is always the possibility that any robot-assisted procedures may be complicated by device malfunction, so it is crucial that the basic skills to perform the procedure manually by the operating surgeon are not lost.

Predictive Analytics

The ability of ML to handle large amounts of data and multiple concurrent variables means it is well suited to elucidating patterns that may not have been revealed by human examination and

statistical methods alone. Machine learning can approach predictive analytics in two main ways, supervised learning and unsupervised learning. In supervised learning an algorithm is derived from a human set of instructions and rules applied to the data. In unsupervised learning the machine is provided with data to process without human “rules” and can therefore identify patterns that may not have already been suspected previously. The most commonly used machine learning models used are support vector machines (SVM) and decision random forests (DRF), these having been associated with the most accurate performances. ML is increasingly being applied to predictive analytics in medicine, from personalized treatment decisions to operational management. In orthopedics ML has been employed for the prediction of surgical outcomes and postoperative complications, with the subspecialty of spinal surgery leading the way in its development and adoption.

Orthopedic Databases

National orthopedic registries, such as the National Joint Registry established in 2002, exist for the measurement of long-term effectiveness and monitoring of safety and patient outcomes [4]. The application of natural language processing, a ML technique, to extract and analyze data from these pioneering databases can provide invaluable insights for the profession. Similar techniques can be applied to smaller hospital databases and EHRs. Though not yet in widespread use, automated algorithms have been developed and trained for this purpose in hip and spine surgery.

Total hip arthroplasty is one of the most reproducible and successful procedures in medicine [3]. Orthopedic bespoke training and validation sets have been used to train NLP algorithms to extract data from orthopedic databases on total hip arthroplasty complications as defined by the American Hip Society. When comparing the output of the orthopedic NLP tool with the manual review of the dataset, the AI performed promisingly with accuracy of 95% versus that of the manual review 94.5%. When the algorithm looked at implant

characteristics, it performed significantly higher than the manual reviewer which unsurprisingly suggests data with definite variables can achieve the highest levels of accuracy. The accuracy for extracting postoperative complications was also higher than the human reviewer, though the algorithm was less accurate at extracting postoperative interventions [56]. This algorithm has been employed locally to develop bespoke consenting for patients based on local data; ML has the potential here to more accurately predict local complications rates for their patients, help surgeons appraise their local practice, and recognize issues early.

The orthopedic subspecialty of spinal surgery has leveraged ML to provide surgeons and patients with accurate and meaningful predictive analytics for use in adult deformity surgery. Spinal deformity, at its most severe, can cause crippling deformity, pain, and disability. Its corrective surgical management is complex and risky but can, if planned carefully, lead to significantly improved health and quality of life for patients. Until recently surgeons relied on personal clinical experience, consensus, and literature based on linear or logistic regression modelling to weigh their decision-making for these complex patients and procedures. More recently, spinal deformity surgeons have incorporated machine learning into their practice. The International Spine Study Group and the European Spine Study Group used prospective multicenter databases to develop an ML-based prognostic tool for major complication, hospital readmission, and unplanned reoperation in adult spinal deformity surgery. The model, created with a random survival forest algorithm, had a predictive accuracy area under curve (AUC) ranging from 0.67 to 0.92 [57]. This has now been developed into a publicly available calculator allowing deformity surgeons to better inform patients on individualized risks from accurate prognostic models and augment their surgical decision-making [58].

Predicting Disease Onset and Degree

If the efficacy of screening for scoliosis improved, this could result in the disease being identified

earlier than clinical assessment alone may allow. In turn, earlier treatment could then result in better outcomes and reduced costs. Measurement of the Cobb angle, which quantifies the magnitude of the spinal deformity, is key for treatment planning in patients with scoliosis. CNNs have been created to identify Cobb angles of greater than 30° with 100% sensitivity and 75% specificity using torso topography and radiographs [59]. ML has also been used to detect and quantify scoliosis, Cobb angle, and spinal alignment using surface topography and moiré images [60].

SVM have been used to quantify curve severity using radiographs, the results of which were statistically comparable to more time-consuming manual measurements [61]. There are multiple iterations of these tools with varying but encouraging levels of accuracy [62].

Postoperative Complications and Rehabilitation

Orthopedic surgical patients often begin their postoperative recovery on orthopedic wards where their recovery may be complicated by bleeding, infection, venous thromboembolism, sepsis, pneumonia, stroke, cardiac arrest, kidney failure, or other life- or limb-threatening post-surgical adverse events.

The analysis of big data presents opportunity to aid clinicians in preoperative decision-making by identifying prognostic factors and risks of postoperative complications and adverse events. An international consensus found that patient comorbid states are more significant drivers of postoperative complications than actual procedural characteristics [63]. Though more research is needed, it would suggest preoperative predictive models (as discussed above) and optimization could significantly help drive down these kinds of adverse events.

Routine vitals monitoring on most postoperative wards is 4 hourly; early decompensations can be missed in the windows between or by infrequent review due to staffing shortages. In the case of many of these complications, taking sepsis as an example, it is clear timely intervention saves

lives [64]. Artificial intelligence applications aimed at prevention or earlier recognition of these postoperative complications can help reduce morbidity, mortality, and length of stay of patients [65, 66]. Wearables data normally marred by data artifact can be mitigated by machine learning algorithms allowing for constant monitoring and earlier warnings prompting assessment. These same technologies can be adopted for community remote monitoring and physical activity functional assessment statuses [67].

Surgical complications such as surgical site infections (SSIs) are of great concern in orthopedic practice. Deep neural networks have been utilized to classify patients at greatest risk of developing this postoperative complication with impressive accuracy.

In spinal surgery, SSI rate is reported ranging 1.2–8.5% increasing morbidity and decreasing patient satisfactory and at significant cost [68]. Deep neural network classification models have been trained to stratify risk factors to provide predictive clinician decision support. One model was able to predict SSI with a positive predictive value of 92.56% and a negative predictive value of 98.45% [69]. These models are currently experimental and not used in widespread practice. More work is necessary to refine such models which suffer limitations including bias and manual alteration of varying model parameters during model construction.

Predictive analytics is an extensive field with endless potential application and is likely to continue to evolve with greater accuracy as further research is conducted and models optimized and further data becomes available.

Conclusion

This chapter has explored a handful of the available AI applications in trauma and orthopedics. Orthopedics is a specialty with a vast patient population and is ripe for the adoption and evolution of AI. AI has multiple applications in orthopedics ranging from detecting occult fractures on imaging to guiding the surgeon's hand to achieve optimal bone cuts intraoperatively. Although application of

technologies such as robotics to orthopaedic practice are still in many cases novel and experimental, they have begun to gain traction as short-term results have become available. With the scale of the specialty, frequency, and reproducibility of procedures, this adoption is likely to rise exponentially as further techniques and data become available.

References

1. Musculoskeletal conditions [Internet]. NHS: Long term conditions. [cited 2021 Feb 28]. <https://www.england.nhs.uk/ourwork/clinical-policy/lte/our-work-on-long-term-conditions/musculoskeletal/>
2. Watkins-Castillo S, Andersson G. United States Bone and Joint Initiative: the burden of musculoskeletal diseases in the United States (BMUS) [Internet]. The Burden of Musculoskeletal diseases in the United States. 2014. <http://www.boneandjointburden.org>
3. Learmonth ID, Young C, Rorabeck C. The operation of the century: total hip replacement. Lancet. 2007;370: 1508–19.
4. NJR Centre. National Joint Registry: Home [Internet]. About. [cited 2021 Mar 3]. <https://www.njrcentre.org.uk/njrcentre/default.aspx>
5. Cabitza F, Locoro A, Banfi G. Machine learning in orthopedics: a literature review. Front Bioeng Biotechnol. 2018;6:75.
6. Karl JW, Swart E, Strauch RJ. Diagnosis of occult scaphoid fractures a cost-effectiveness analysis. J Bone Jt Surg Am. 2014;97(22):1860–8.
7. Lau S, Bozin M, Thillainadesan T. Lisfranc fracture dislocation: a review of a commonly missed injury of the midfoot. Emerg Med J. 2017;34(1):52–6.
8. Pinto A, Berritto D, Russo A, Ricciutello F, Caruso M, Belfiore MP, et al. Traumatic fractures in adults: missed diagnosis on plain radiographs in the emergency department. Acta Biomed. 2018;89:111–23.
9. Clementson M, Björkman A, Thomsen NOB. Acute scaphoid fractures: guidelines for diagnosis and treatment. EFORT Open Rev. 2020;5(2):96–103.
10. Nunley JA, Vertullo CJ. Classification, investigation, and management of midfoot sprains: Lisfranc injuries in the athlete. Am J Sports Med. 2002;30(6):871–8.
11. Lee YH. Efficiency improvement in a busy radiology practice: determination of musculoskeletal magnetic resonance imaging protocol using deep-learning convolutional neural networks. J Digit Imaging. 2018;31(5):604–10.
12. Trivedi H, Mesterhazy J, Laguna B, Vu T, Sohn JH. Automatic determination of the need for intravenous contrast in musculoskeletal MRI examinations using IBM Watson's natural language processing algorithm. J Digit Imaging. 2018;31(02):245–51.
13. Dutta S, Long WJ, Brown DFM, Reisner AT. Automated detection using natural language

- processing of radiologists recommendations for additional imaging of incidental findings. *Ann Emerg Med.* 2013 Aug;62(2):162–9.
14. Kunze KN, Rossi DM, White GM, Karhade AV, Deng J, Williams BT, et al. Diagnostic performance of artificial intelligence for detection of anterior cruciate ligament and meniscus tears: a systematic review. *Arthrosc J Arthrosc Relat Surg [Internet].* 2021;37(2): 771–81. <https://doi.org/10.1016/j.arthro.2020.09.012>.
15. Diermeier T, Rothrauff BB, Engebretsen L, Lynch AD, Ayeni OR, Paterno MV, et al. Treatment after anterior cruciate ligament injury: Panther Symposium ACL Treatment Consensus Group. *Knee Surg Sports Traumatol Arthrosc [Internet].* 2020;28(8):2390–402. <https://doi.org/10.1007/s00167-020-06012-6>.
16. Stone JA, Perrone GS, Nezwak TA, Cui Q, Vlad SC, Richmond JC, et al. Delayed ACL reconstruction in patients \geq 40 years of age is associated with increased risk of medial meniscal injury at 1 year. *Am J Sports Med.* 2019;47(3):584–9.
17. Hammernik K, Klatzer T, Kobler E, Recht MP, Sodickson DK, Pock T, et al. Learning a variational network for reconstruction of accelerated MRI Data. *arXiv.* 2017.
18. Hirschmann A, Cyriac J, Stieltjes B, Kober T, Richiardi J, Omoumi P. Artificial intelligence in musculoskeletal imaging: review of current literature, challenges, and trends. *Semin Musculoskelet Radiol.* 2019;23(3):304–11.
19. Olczak J, Fahlberg N, Maki A, Razavian AS, Jilert A, Stark A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs: deep learning algorithms – are they on par with humans for diagnosing fractures? *Acta Orthop.* 2017;88(6):581–6.
20. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clin Radiol [Internet].* 2018;73(5):439–45. <https://doi.org/10.1016/j.crad.2017.11.015>.
21. Bien N, Rajpurkar P, Ball RL, Irvin J, Park A, Jones E, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. Saria S, editor. *PLOS Med [Internet].* 2018 Nov 27 [cited 2021 Mar 3];15(11):e1002699. <https://dx.plos.org/10.1371/journal.pmed.1002699>
22. Štajduhar I, Mamula M, Miletic DÜG. Semi-automated detection of anterior cruciate ligament injury from MRI. *Comput Methods Prog Biomed.* 2017;140:151–64.
23. Chang PD, Wong TT, Rasiej MJ. Deep learning for detection of complete anterior cruciate ligament tear. *J Digit Imaging.* 2019;32(6):980–6.
24. Garwood ER, Tai R, Joshi G, Watts VGJ. The use of artificial intelligence in the evaluation of knee pathology. *Semin Musculoskelet Radiol.* 2020;24(1):21–9.
25. Gorelik N, Chong J, Lin DJ. Pattern recognition in musculoskeletal imaging using artificial intelligence. *Semin Musculoskelet Radiol.* 2020;24(1):38–49.
26. Norman B, Pedoia V, Noworolski A, Link TM, Majumdar S. Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs. *J Digit Imaging.* 2019;32(3):471–7.
27. Quatman CE, Hettrich CM, Schmitt LC, Spindler KP. The clinical utility and diagnostic performance of magnetic resonance imaging for identification of early and advanced knee osteoarthritis: a systematic review. *Am J Sports Med.* 2011;39(7):1557–68.
28. Wilson NA, Juhn M, York S, Davis CM. Revision total hip and knee arthroplasty implant identification: implications for use of unique device identification 2012 AAHKS member survey results. *J Arthroplast.* 2014;29(2):251–5.
29. Karnuta JM, Haebel HS, Luu BC, Roth AL, Molloy RM, Nystrom LM, et al. Artificial intelligence to identify arthroplasty implants from radiographs of the hip. *J Arthroplasty [Internet].* 2020; <https://doi.org/10.1016/jarth.2020.11.015>.
30. Borjali A, Chen AF, Muratoglu OK, Morid MA, Varadarajan KM. Detecting total hip replacement prosthesis design on preoperative radiographs using deep convolutional neural network. *arXiv.* 2019;1–16.
31. Lodwick G, Haun C, Smith W, Keller R, Robertson E. Computer diagnosis of primary bone tumors: a preliminary report. *Radiology.* 1963;80(2):273–5.
32. Ping YY, Yin CW, Kok LP. Computer aided bone tumor detection and classification using x-ray images. *IFMBE Proc.* 2008;21 IFMBE(1):544–7.
33. Bandyopadhyay O, Biswas A, Bhattacharya BB. Bone-cancer assessment and destruction pattern analysis in long-bone X-ray image. *J Digit Imaging.* 2019;32(2):300–13.
34. Suhas MV, Mishra A. Classification of benign and malignant bone lesions on CT images using random forest. In: 2016 IEEE international conference on recent trends in electronics, information and communication technology, RTEICT 2016 – proceedings. 2017. p. 1807–10.
35. Lang JE, Mannava S, Floyd AJ, Goddard MS, Smith BP, Mofidi A, et al. Robotic systems in orthopaedic surgery. *J Bone Jt Surg Ser B.* 2011;93 B(10): 1296–9.
36. Mavrogenis AF, Scarlat MM. Surgeons and robots. *Int Orthop.* 2019;43(6):1279–81.
37. Huang H-M, Messina E, Albus J. Autonomy levels for unmanned systems (ALFUS) framework volume II: framework models version 1.0. 2007.
38. Picard F, Deakin AH, Riches PE, Deep K, Baines J. Computer assisted orthopaedic surgery: past, present and future. *Med Eng Phys.* 2019;72:55–65.
39. Han X, Tian W, Liu Y, Liu B, He D, Sun Y, et al. Safety and accuracy of robot-assisted versus fluoroscopy-assisted pedicle screw insertion in thoracolumbar spinal surgery: a prospective randomized controlled trial. *J Neurosurg Spine.* 2019;30(5):615–22.
40. Davies BL, Rodriguez Y, Baena FM, Barrett ARW, Gomes MPSF, Harris SJ, Jakopec M, et al. Robotic

- control in knee joint replacement surgery. *Proc Inst Mech Eng H.* 2007;221(1):71.
41. Batailler C, Swan J, Marinier ES, Servien EL, Lustig S. New technologies in knee arthroplasty: current concepts. *J Clin Med.* 2021;10:47.
 42. Pearle AD, O'loughlin PF, Kendoff DO. Robot-assisted unicompartmental knee arthroplasty. *J Arthroplast.* 2010;25(2):230.
 43. St Mart JP, De Steiger RN, Cuthbert A, Donnelly W. The three-year survivorship of robotically assisted versus non-robotically assisted unicompartmental knee arthroplasty. *Bone Jt J.* 2020;102 B(3):319–28.
 44. Plate JF, Mofidi A, Mannava S, Smith BP, Lang JE, Poehling GG, et al. Achieving accurate ligament balancing using robotic-assisted unicompartmental knee arthroplasty. *Adv Orthop.* 2013;2013:1–6.
 45. Chen AF, Kazarian GS, Jessop GW, Makhdum A. Current concepts review: robotic technology in orthopaedic surgery. *J Bone Jt Surg Am.* 2018;100(22):1984–92.
 46. Kayani B, Konan S, Thakrar RR, Huq SS, Haddad FS. Assuring the long-term total joint arthroplasty: a triad of variables. *Bone Jt J.* 2019;101B(1):11–8.
 47. Choong PF, Dowsey MM, Stoney JD. Does accurate anatomical alignment result in better function and quality of life? Comparing conventional and computer-assisted total knee arthroplasty. *J Arthroplast.* 2009;24(4):560–9.
 48. Hiscox CM, Bohm ER, Turgeon TR, Hedden DR, Burnell CD. Randomized trial of computer-assisted knee arthroplasty: impact on clinical and radiographic outcomes. *J Arthroplast.* 2011;26(8):1259–64.
 49. Christ AB, Pearle AD, Mayman DJ, Haas SB. Robotic-assisted unicompartmental knee arthroplasty: state-of-the art and review of the literature. *J Arthroplast.* 2018;33(7):1994–2001.
 50. Kayani B, Konan S, Pietrzak JRT, Huq SS, Tahmassebi J, Haddad FS. The learning curve associated with robotic-arm assisted unicompartmental knee arthroplasty. *Bone Jt J.* 2018;100B(8):1033–42.
 51. Jassim SS, Benjamin-Laing H, Douglas SL, Haddad FS. Robotic and navigation systems in orthopaedic surgery: how much do our patients understand? *CiOs Clin Orthop Surg.* 2014;6(4):642.
 52. Kim Y-H, Yoon S-H, Park J-W. Does robotic-assisted TKA result in better outcome scores or long-term survivorship than conventional TKA? A randomized, controlled trial. *Clin Orthop Relat Res [Internet].* 2020 Feb 1 [cited 2021 Mar 4];478(2):266–75. <https://journals.lww.com/10.1097/CORR.0000000000000916>
 53. Illgen RL, Bukowski BR, Abiola R, Anderson P, Chughtai M, Khlopas A, et al. Robotic-assisted total hip arthroplasty: outcomes at minimum two-year follow-up. *Surg Technol Int.* 2017;30:365–72.
 54. Bukowski BR, Anderson P, Khlopas A, Chughtai M, Mont MA, Illgen RL. Improved functional outcomes with robotic compared with manual total hip arthroplasty. *Surg Technol Int.* 2016;29:303–8.
 55. Haddad FS, Horriat S. Robotic and other enhanced technologies: are we prepared for such innovation? *Bone Jt J.* 2019;101-B(12):1469–71.
 56. Wyles CC, Tibbo ME, Fu S, Wang Y, Sohn S, Kremers WK, et al. Use of natural language processing algorithms to identify common data elements in operative notes for total hip arthroplasty. *J Bone Jt Surg Am [Internet].* 2019 Nov 6 [cited 2021 Mar 3];101(21):1931–8. <https://pubmed.ncbi.nlm.nih.gov/31567670/>
 57. Pellisé F, Serra-Burriel M, Smith JS, Haddad S, Kelly MP, Vila-Casademunt A, et al. Development and validation of risk stratification models for adult spinal deformity surgery. *J Neurosurg Spine [Internet].* 2019 Oct 1 [cited 2021 Mar 3];31(4):587–99. <https://thejns.org/spine/view/journals/j-neurosurg-spine/31/4/article-p587.xml>
 58. ESSG|Research Projects [Internet]. [cited 2021 Mar 3]. <http://www.spine-essg.com/web/research-projects/research-awards/>
 59. Jaremko JL, Poncet P, Ronsky J, Harder J, Dansereau J, Labelle H, et al. Estimation of spinal deformity in scoliosis from torso surface cross sections. *Spine (Phila Pa 1976).* 2001;26(14):1583–91.
 60. Watanabe K, Aoki Y, Matsumoto M. An application of artificial intelligence to diagnostic imaging of spine disease: estimating spinal alignment from Moiré images. *Neurospine.* 2019;16:697–702.
 61. Duong L, Cheriet F, Labelle H. Automatic detection of scoliotic curves in posteroanterior radiographs. *IEEE Trans Biomed Eng.* 2010;57(5):1143–51.
 62. Chen K, Zhai X, Sun K, Wang H, Yang C, Li M. A narrative review of machine learning as promising revolution in clinical practice of scoliosis. *Ann Transl Med [Internet].* 2021 [cited 2021 Mar 3];9(1):67. <https://doi.org/10.21037/atm-20-5495>.
 63. Hassen YAM, Johnston MJ, Singh P, Pucher PH, Darzi A. Key components of the safe surgical ward. *Ann Surg [Internet].* 2019 Jun 1 [cited 2021 Mar 2];269(6):1064–72. <https://journals.lww.com/00000658-201906000-00011>
 64. NHS England. Factsheet: implementation of the “Sepsis Six” care bundle. 2014;(February):2013–5. <https://www.england.nhs.uk/wp-content/uploads/2014/02/rm-fs-10-1.pdf>
 65. Hellings TS, Martin LC, Martin M, Mitchell ME. Failure events in transition of care for surgical patients. *J Am Coll Surg.* 2014;218(4):723–31.
 66. Sun EC, Darnall BD, Baker LC, MacKey S. Incidence of and risk factors for chronic opioid use among opioid-naïve patients in the postoperative period. *JAMA Intern Med [Internet].* 2016 Sep 1 [cited 2021 Mar 3];176(9):1286–93. <https://jamanetwork.com/>
 67. Loftus TJ, Tighe PJ, Filiberto AC, Balch J, Upchurch GR, Rashidi P, et al. Opportunities for machine learning to improve surgical ward safety. *Am J Surg [Internet].* 2020;220(4):905–13. <https://doi.org/10.1016/j.amjsurg.2020.02.037>.
 68. Barker FG. Efficacy of prophylactic antibiotic therapy in spinal surgery: a meta-analysis. *Neurosurgery*

- [Internet]. 2002 Aug 1 [cited 2021 Mar 3];51(2):391–401. <https://academic.oup.com/neurosurgery/article/2739794/Efficacy>
69. Hopkins BS, Mazmudar A, Driscoll C, Svet M, Goergen J, Kelsten M, et al. Using artificial intelligence (AI) to predict postoperative surgical site infection: a retrospective cohort of 4046 posterior spinal fusions. Clin Neurol Neurosurg [Internet]. 2020;192(December 2019):105718. <https://doi.org/10.1016/j.clineuro.2020.105718>.



Harnessing Artificial Intelligence in Maxillofacial Surgery

64

Karishma Rosann Pereira

Contents

Introduction and Background	888
The Maxillofacial Surgeon and AI	889
Is AI a Friend or Foe?	889
Simplifying AI for the Surgeon: Suggestions for Seamless Integration of AI and Surgery	890
Machine Learning and Deep Learning	890
Artificial Neural Networks	891
Natural Language Processing	893
Computer Vision	894
Surgeon Dilemmas on AI	895
Role of Surgeons in Enabling AI-Assisted Maxillofacial Surgeries	897
Challenges in the Path	898
Suggestions for Overcoming This Challenge	898
Suggestions for Overcoming This Challenge	898
Suggestions for Overcoming This Challenge	899
Literature Speaks: An Overview of Current Applications with Potential for Harnessing AI in Maxillofacial Surgery	899
Maxillofacial Presurgical Imaging	899
Orthognathic Surgery	899
Implant Surgery	900
Temporomandibular Joint (TMJ) Surgery	900
Oncosurgery and Reconstruction	901
Trauma Surgery	901
Impacted Teeth and Minor Oral Surgery	902
Miscellaneous	902
Watch List for Future Forward Maxillofacial Surgeons	902
Surgical Data Science	902
Surgical Scene Analytics	903
Surgical Control Tower (SCT)	903

K. R. Pereira (✉)
KNR University of Health Sciences, Hyderabad, India

Conclusion	903
References	903

Abstract

Indeed we live in exciting times, at the cusp of a world which is a melting pot of novel technologies and surgical advances. Artificial intelligence holds mind-boggling potential for the informed surgeon who is open to thinking out of the box and expanding learning horizons. Boundaries must merge; interdisciplinary teamwork must emerge to propel a healthcare revolution. This chapter heralds a new age in surgery, where the power of machine learning meets the surgeon's clinical acumen. The humble healing scalpel becomes a hi-tech widget! The ultimate aim is healthcare that is affordable, accessible, and par excellence so as to improve patient quality of life and disease combat.

Keywords

Artificial intelligence · Deep learning · Machine learning · Maxillofacial surgery · Natural language processing · Neural networks · Oncosurgery · Orthognathic surgery · Temporomandibular joint disorders

Introduction and Background

Tracing the identity of oral and maxillofacial surgery leads to a deeper understanding of its origin, scope, and emerging roles. This branch of surgery has come a long way from the times of historically pioneering surgeons such as Simon P. Hullihen and James E. Garretson who named this branch and is considered the founder of oral surgery. The specialty deals with conditions and defects, both congenital and acquired, of the head, face, oral cavity, and neck. The maxillofacial surgeon of today needs to be equipped with training in delicate and masterful surgical techniques on prominent organs involved in vital body functions. From trauma and pathology to orthognathic, esthetic, cleft lip/palate, the maxillofacial trainee

of today must be abreast with complex lifesaving oncosurgical resections, reconstructive procedures including the use of distant free flaps, and technology-enhanced surgical care (Images 1 and 2).

Healthcare is witnessing an emerging domain of artificial intelligence (AI). A concept that seemed out of a science fiction movie is now an accessible reality. It is therefore the need of the hour to wake up to this compelling field and equip medical practitioners with the principles and practical knowledge on the subject. The convergence of medical professionals and their skill set and the potential of AI will herald the new era of "human-tech-driven intelligence."

Artificial intelligence surfaced amidst awe as well as alarm. As with any novel disruptive concept, it has been through the "hype cycle." Despite controversies, AI has been increasingly finding applications in the world around us, and naturally healthcare too has embraced AI.

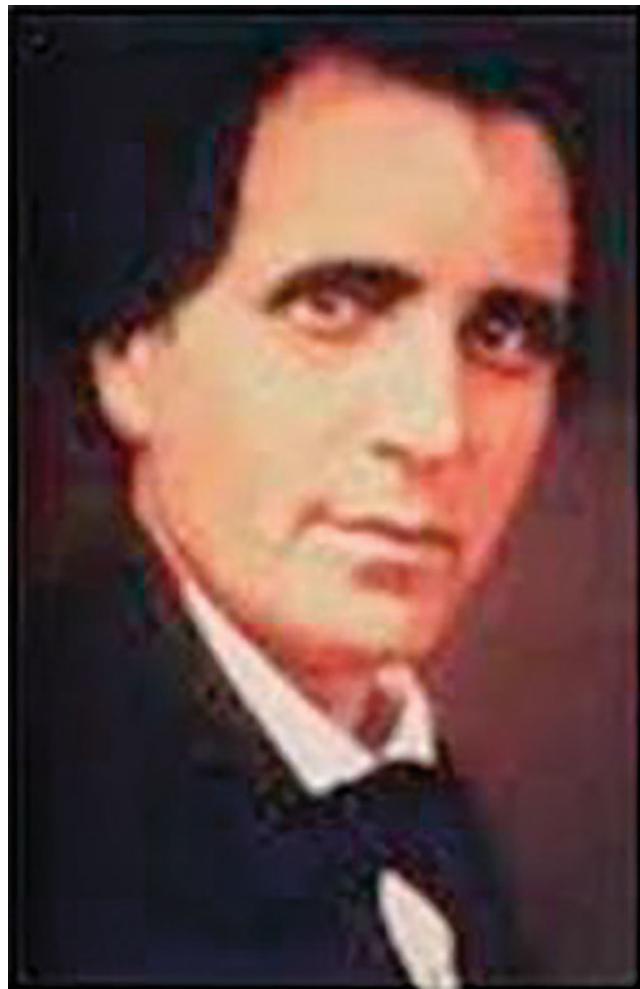
The role of AI in medicine has expanded centripetally. AI gradually inched its way from outpatient data analytics that transcended into better streamlined patient scheduling, doctor workflow optimization, resource management, etc. AI then proved its merits in radiological and pathological diagnostics and enhanced predictive capacity of clinicians. AI in the OT (operation theatre) is the ultimate goal of a futuristic surgeon.

In order to be in a position to use and develop AI tools in maxillofacial surgery, it is imperative that every surgeon holds a basic understanding of the fundamentals of this science.

This chapter aims to

- Serve as a ready reckoner on AI for maxillofacial surgeons.
- Bust myths and clarify the topic to surgeons who are not trained engineers or tech gurus.
- Shine light on principles and applications of AI by means of literature review.
- Lay the foundation and provide the insight upon which the readers can then build on and

Image 1 Dr. Simon P. Hullihen. <http://accessomfs.com/about> or Ref. [39]



develop unique tools in subspecialties of maxillofacial surgery.

- (E) Help practicing surgeons develop a thought process in the field of AI.

solving. Computer science defines AI as “intelligent agents” capable of perceiving their environment and more importantly taking learned actions to maximize their abilities of successfully executing a planned/programmed goal [1].

However, the very word “artificial” indicates an intelligence that is *unlike natural* intelligence inherent in humans. The chief distinguishing factors between the two are consciousness and emotion. Key aspects of human intellectual acumen are intuition and empathy, components AI will most likely struggle to mimic.

The current consensus from literature on AI in the healthcare points toward improved outcomes by joint human-machine tasking rather than superiority of either entity over the other [2]. Looking back we may recall how computers gained

The Maxillofacial Surgeon and AI

Is AI a Friend or Foe?

To get to the bottom of this dilemma, let us ascertain what AI definitively is. Simply put, AI can be described as machines that exhibit human-like intelligence. It involves the use of “algorithms” that give machines the capacity to reason and perform functions such as word, speech, and object recognition, decision-making, and problem-

Image 2 Dr. James E. Garretson. <http://accessomfs.com/about> or <https://www.jstor.org/stable/44445732?seq=1>



popularity. Then came the era of early Internet followed by mobile phones that today have evolved into smart phones without which we can hardly imagine our day go by. The anticipated benefits and unanticipated risks associated with any new applied science will also exist in the case of AI. In the medical field, there would be additional concerns of ethical and medicolegal implications to be sorted out.

Ultimately the bottom line is that AI can be a friend if humans intelligently harness its potential. Awareness of possible pitfalls is essential to preventing AI from becoming our foe.

Simplifying AI for the Surgeon: Suggestions for Seamless Integration of AI and Surgery

Data is the starting point, and surgeons serve as the most essential “data providers” of this

cornerstone requisite. The surgeon is the point of contact with the patient. The surgeon’s training, knowledge of disease pathophysiology, pharmacology, complications, management, and expertise of surgical techniques are what should be developed into a data bank. This data will then serve as fodder to develop algorithms that are the building blocks of the AI chain. This would require for all practical purposes, a collaboration between surgeons and AI engineers.

The various subsets of AI include machine learning (ML), deep learning (DL), artificial neural networks (ANN), natural language processing (NLP), and computer vision to name a few.

Machine Learning and Deep Learning

The machine analyzes data sets and detects sequences and patterns to come up with predictions. While machine learning strings together

data into algorithms from which it learns and makes predictions, deep learning strings together several layers of algorithms into artificial neural networks which are capable of making informed decisions on their own.

Three broad processes by which machine learning occurs are the following:

- (A) Supervised Learning – Data that has been labeled by a human is filed into a machine learning algorithm. The aim is to teach the computer a function such as recognizing the fracture line in a maxillofacial skeletal radiograph or detecting a complication post-surgery based on multiple case sheet data.
- (B) Unsupervised Learning – The machine learning algorithm receives unlabeled data. The machine itself will then find a hidden pattern within the data, for example, identifying various tissues based on color or density – the bone, blood, mucosa, skin, etc.
- (C) Reinforcement Learning – An advanced form of ML similar to operant conditioning (which is a form of associative learning in which a desired behavior is established by modification via reinforcement of good and punishment of bad behavior). Here the machine itself learns from its mistakes and success. It is particularly useful in developing automated surgical systems.

Suggested Applications of ML and DL in Maxillofacial Surgery

- AI drug delivery bots in chronic stubborn maxillofacial infections such as osteomyelitis, osteoradionecrosis, and cancerum oris. The system would recognize patterns within the wound environment variables such as type of causative organism (gram positive/negative, species, aerobic/anaerobic) microbial counts, rate of neoangiogenesis, fibroblast count and collagen synthesis turnover, and leukocyte counts (factors which are indiscernible to the human eye). The machine can learn to establish patterns such as -at what level of microbial contamination what symptom is elicited and can predict surgical site infection. Furthermore, when many algorithms work together (ensemble ML), the machine can use these generated patterns to access medical literature

and arrive at the best medication and administer timely dose modifications based on real-time changes in the wound microenvironment, thus eliminating the time delay in traditional laboratory tests and solving the pressing universal issue of the increasing ill effects of antibiotic resistance.

- Self-adjusting distraction osteogenesis (DO) devices that would identify bone morphology and histology patterns and self-formulate the ideal Ilizarov principles to ensure surgical success. The AI system would enlist the optimal latency period, distraction rate and rhythm, and consolidation phase along with providing the surgeon the best osteotomy site and plan with minimal periosteal stripping for perfect callus and new bone in cases of hypoplastic jaws and hemifacial microsomia (Images 3 and 4).

These are just two probable applications presented. Several uses of machine learning in maxillofacial surgery are awaiting to be developed, and hopefully the above examples will have served as inspiration for the same.

Artificial Neural Networks

Probably the most interesting branch of AI, made of up several “processing elements (PEs). There is an input level/layer, middle multiple hidden levels/layers, and an output level/layer. Data is received at the input level. Calculations, summations, learning, and analysis of complex nonlinear relationships between dependent and independent data variables all occur in the many hidden layers. The hidden layers finally send the determined prediction for interpretation to the output layer [3, 4].

Therefore, the applications of ANN can be vast, and surgeons come into the picture as subject matter experts to provide a knowledge base to be fed into these algorithms that form the ANNs.

The potential reach of ANNs in maxillofacial surgery range from diagnosis and prognosis [5, 6], classifications and predictions [7–10], pattern recognition [11], controlled drug design and delivery systems [12], and much more.

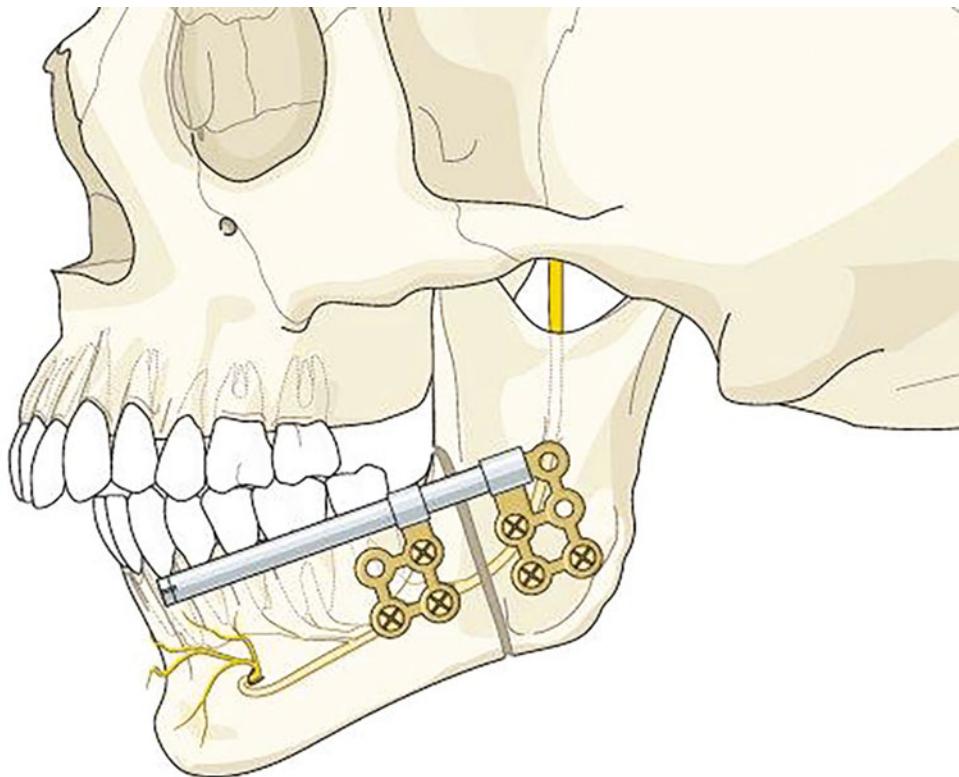


Image 3 Distraction osteogenesis of the mandible.
https://www.researchgate.net/publication/326961118_Distraction_Osteogenesis_in_Oral_and_Maxillofacial_Reconstruction_Applications_Feasibility_Study_of_Design_and_Development_of_an_Automatic_Continous_Distractor/citations

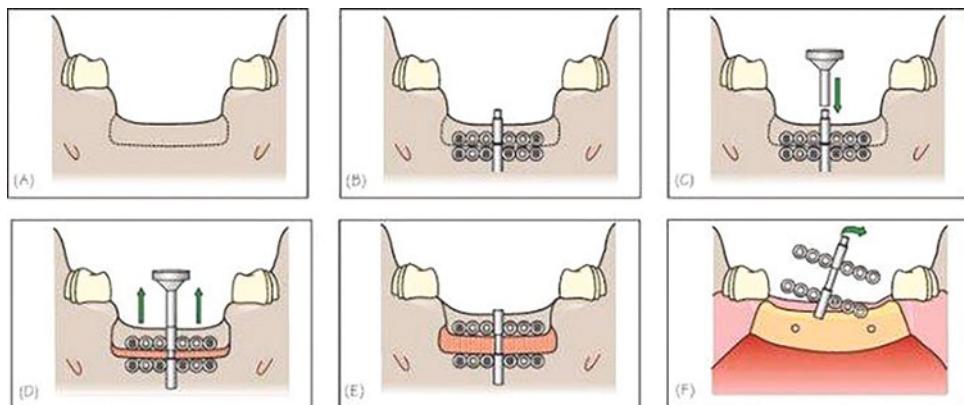
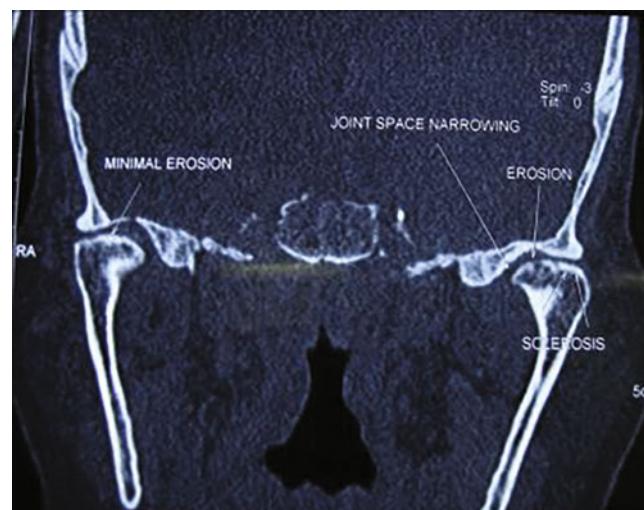


Image 4 Distraction osteogenesis procedure.
https://www.researchgate.net/publication/326961118_Distraction_Osteogenesis_in_Oral_and_Maxillofacial_Reconstruction_Applications_Feasibility_Study_of_Design_and_Development_of_an_Automatic_Continous_Distractor/citations

Suggested Applications of ANNs in Maxillofacial Surgery

- In the early diagnosis of oral cancer lesions and early onset osteoarthritic changes in the temporomandibular joints. These are two areas where early diagnosis can save life and reduce tremendous burden of complicated surgeries if the disease is caught at an advanced stage. The quest for noninvasive and alternative diagnostic aids like salivary biomarkers has been on for oral cancer screening, but clinical trials can take years to translate to full-scale use. On the other hand, ANNs can be successfully trained to detect and predict clinical outcomes of node-positive oral cancer patients. This is possible by feeding the network with input information of past patients such as patients' habit history, age, tumor size, number of cervical lymph nodes, DNA index and S-phase determination by flow cytometry, clinical follow-up, relapse rate and time, etc. (Image 5).
- ANNs can be constructed based on standard radiographic images and histological specimens' data and then trained to distinguish benign from malignant lesions. Thus, the positive predictive value of biopsies can tremendously be enhanced.
- Development of ANNs to detect hemodynamic, thermal, physical changes in postoperative ICU patients and attached to alarm

Image 5 AI trained with CT images of TMJs showing osteoarthritic changes [41]



systems will be a revolution in critical postsurgical care.

- ANNs can herald an era of smart operating room monitoring of the Boyles apparatus and patient vitals intraoperatively.

Natural Language Processing

Training a computer to comprehend human language and furthermore infer meaning from vast and unstructured data such as case files, operative notes, and clinical findings. Gradually in time, the system is able to recognize complex sentences and thus enables surgeons to input their notes more naturally rather than being confined to specific drop-down menus or computer code words.

This automation can impact millions in positive ways through downloadable apps that can translate in several world languages.

Suggested Applications of NLP in Maxillofacial Surgery

- Interhospital communication using NLP systems to compare perioperative parameters, complications, and management protocols and arrive as standardized and optimized benchmarks.
- Virtual OT and ICU assistant and scribe can be a game changer. Surgeons can dictate notes that would be efficiently documented with

higher accuracy and retrievability. These notes would further enhance the system's ability to teach itself for future use. The NLP would then be able to comb through all the progress notes and operative reports to deduce predictions that would aid other patients.

Computer Vision

This subset refers to object recognition by systems that have learned to understand images and video data. Surgery unlike other specialties of medicine and diagnostics is heavily involved with dynamic and dexterous movement-based procedures. In surgery, AI has to analyze a dynamic environment data as opposed to static images. To be of effective use during an operation, AI must be able to deliver real-time solutions. This is where computer vision along with deep learning and other subsets of AI will be of particular significance. Though predictive video AI analytics is yet in very initial stages, it holds great potential. It is estimated that even a high-resolution computed tomographic (CT) image contains 25 times less data value than a mere 1-min high-definition surgical video clip.

Thus, one can imagine the tremendous use of computer vision if leveraged to process surgical videos and images to identify and predict adverse events in real time for intraoperative assistance [13].

What is being discussed here is not to be confused with the existing laparoscopic, minimally invasive, and image-guided robotic systems already in use like the TORS (transoral robotic surgery) for oncological resections and arthroscopic surgeries of the temporomandibular joint.

Suggested Applications of Computer Vision in Maxillofacial Surgery

- Computer vision based on deep learning to improve the ability of the system to reliably automate nuances of suturing actions (AI autonomous suturing) in an uncontrolled and live anatomical scenario. In time, the system understands from previous suturing videos and is able to decide upon an apt suture (simple continuous, vertical/horizontal mattress) for a particular surgical site and perform it accurately ensuring adequate tension, closure, and precision (Image 6).



Image 6 AI learned suturing by studying surgical videos. <https://thenextweb.com/neural/2020/06/23/how-an-ai-learned-to-stitch-up-patients-by-studying-surgical-videos/>

- Computer vision assessment and decision of third molar surgical extractions based on analysis of ramal/tuberosity/alveolar bone cover, bone texture, and inferior alveolar and lingual nerve proximity. Ability to integrate this information with deep learning to provide the surgeon with best access path and flap options and predict patient risks based on medical history analytics (Image 7).
- Facial esthetic procedures are poised to benefit the most with computer vision AI analytical tools. Orthognathic facial measurements and surgeries will become ultraprecise as the human eye is no compare with computer vision when it comes to minute millimeter assessments.
- Prediction of duration of surgery by analysis of previous surgeries will allow estimation of anesthetics and postsurgical care.
- Computer vision along with augmented reality is already in nascent stages and will soon be the norm in education and training of surgical students. Universities can provide the most

realistic simulation of the human body by means of full-body robotics and augmented reality projected images of the surgical field (Image 8).

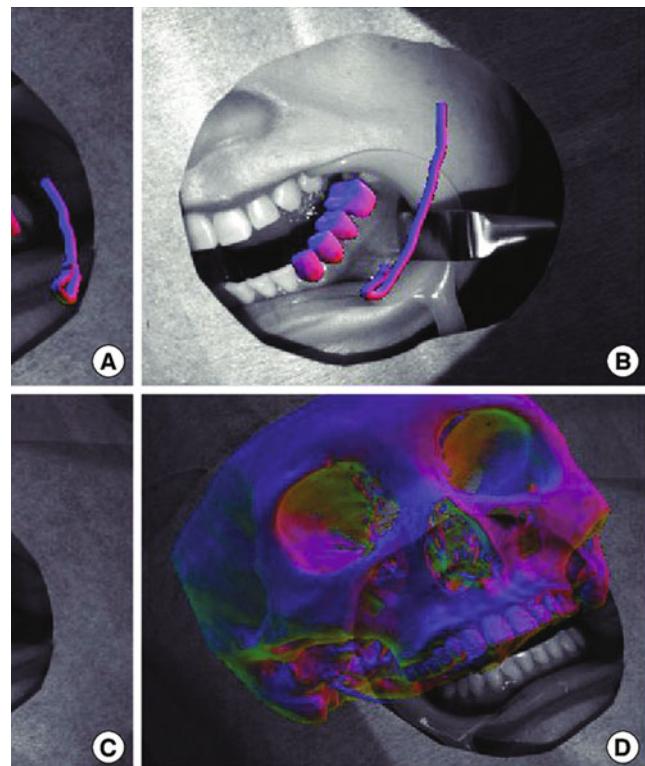
- Planning and performing mock surgery on the CT-derived AI-projected three-dimensional patient's maxillofacial region.

Surgeon Dilemmas on AI

1. Will AI take over my job?

360° Perspective – There have been strong opponents who have opined that AI would displace people from jobs and man is in a race with machine. These claims resonated with the public during periods of economic recession and meltdown. However, this argument fails to support the fact that AI like any other automation system is built to increase speed and accuracy and cannot mimic human multifaceted faculties and senses. An important point to note here is that AI is highly

Image 7 Augmented fusion of patient models for surgical field visualization (up, inferior alveolar nerve; down, maxillofacial skeleton). https://www.researchgate.net/publication/317211457_Virtual_Reality_and_Augmented_Reality_in_Plastic_Surgery_A_Review/figures or Adapted from Wang et al. Int J Med Robot 2016;2016 Jun 9 [Epub]. <https://doi.org/10.1002/rcs.1754> [23], with permission of John Wiley and Sons via Copyright Clearance Center



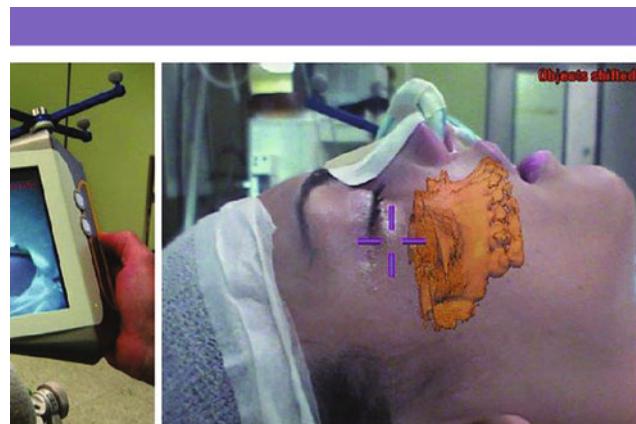


Image 8 Augmented reality projected surgical site (maxilla in orthognathic surgery). https://www.researchgate.net/publication/317211457_Virtual_Reality_and_Augmented_Reality_in_Plastic_Surgery_A_Review/figures?lo=1

or Adapted from Mischkowski et al. J Craniomaxillofac Surg 2006;34:478–83 [21], with permission of Elsevier via Copyright Clearance Center

specific. It can be trained to perform one/few specific tasks, for example, determining whether a premalignant lesion is showing dysplasia or malignant transformation. Even if the AI performs this task better than the surgeon, the fact remains that a surgeon does much more than that. The surgeon communicated with the patient and co-workers, interprets the findings, and determines the type of biopsy, surgical access, armamentarium, type of surgery, medication combinations, postoperative care, and follow-up. One would require several different AI tools to be able to perform all these tens of tasks that a human workforce is capable of doing. AI could enhance accuracy of certain activities within an occupation but is highly unlikely to replace the entire occupation. Would a person experiencing persistent clicking sounds in his jaw use his smart phone to diagnose or head to his physician for an examination? The physician could then augment his clinical examination with an AI tool to improve the delivered care. Thus, AI will augment skill and expand output rather than replace workers. Therefore, the verdict from literature review, history, and plain logic is that any technology that enhances productivity leads to growth of a sector and not job cuts. Furthermore, feeling threatened by technology is not the way forward; instead helping people

understand and receive training in the new methods will give them a sense of empowerment instead of fear.

2. Will AI reduce my critical thinking, diagnostic skill, and clinical expertise?

360° Perspective – On the contrary, AI will help you make smarter decisions, heighten your diagnostic ability, and lead to evolved clinical skill set. Think of it as something that complements your surgical expertise and supports you rather than a competitor. For example, humans may have forgotten how to read paper maps with the advent of GPS-guided navigation systems, but their skill set has evolved to include Internet usage and computer navigation. Thus, AI is only reaffirming the fact that change is the only constant. It is a matter of perspective. AI will reduce laborious repetitive tasks and provide the surgeon with time to spare on tasks requiring innovation, thought, and insight. In high patient load hospitals and limited resource facilities where the patient-to-surgeon ratio is high, AI will prove highly beneficial, as it would finish mundane tasks, thus enabling the surgeon to have more time to concentrate on roles like direct patient consultations, clinical examinations, and higher cognitive activities. Remote areas and small town hospitals where access to specialist care is unavailable can leverage AI decision

systems. AI tools if smart can be used by smarter humans to learn from and improve competence.

3. *But AI could be prone to errors, isn't that a disaster?*

360° Perspective – Who/what is error-free? While on the topic of errors, one can't help but recall the famous adage "To err is Human." An automated system eliminates human errors. Machines do not tire and can perform repetitive tasks infinitely without being prone to negligence. While speculative possibility of errors in AI decision-making does exist, the point remains that on an average the errors computational machinery can make would be less than a human.

4. *We are in for trouble as AI is a technology that makes computers like human brains.*

360° Perspective – That is a notion propagated by fantasy sci-fi movies. In reality, AI only takes inspiration in naming its functional units after structural components of the human nervous system. There is no physiological resemblance, and it would be a mythical distant dream to conjure images of human-like computers. Existing scientific literature is not in favor of imagining machines that mimic the complex functioning, interconnectivity, and biofeedback loops between human neurons. Furthermore, AI is not a technology rather a methodology. AI uses mathematical methods and algorithms to analyze data and adjust parameters via learning rules.

Role of Surgeons in Enabling AI-Assisted Maxillofacial Surgeries

It is high time surgeons assume a role of leadership and step up from being mere spectators of this science to being key drivers of an AI integrated surgical era. The key element is the strategic working advantage that a surgeon holds, being the point of contact with patients. This is a two-way mantle: (A) serving as data input sources to train AI systems and (B) serving as surgical AI conveyors to the patients.

Initial steps include

- (A) Getting acquainted with the basics of AI.
- (B) A crucial move which would require involvement of hospital administrations and novel hospital management policies to develop new departments within hospitals such as medical data analytics and health informatics/healthcare technology. This integrating and partnering with data scientists and engineers is essential to make AI surgical systems of real value. Considering the routine of a surgeon who is almost always scrubbed in an OT uniform and rotates between ward rounds, clinics, ICU, and operating rooms, it would be impractical to imagine surgeons heading to meet AI engineers outside the workspace and vice-versa. It would therefore be advisable to have all concerned professionals work under the same roof. Only then can inter-specialty understanding and knowledge cross-flow develop. Surgeons can impart their knowledge on complex anatomy, etiopathogenesis, pharmacology, disease clinical course, and surgical techniques to the AI engineering team while understanding the development needs of the technology team. This would be the fastest way to bring AI to the forefront of maxillofacial surgery.
- (C) A collaborative effort on the part of surgeons is required to form a surgical pool similar to biobanks and genomics. This could facilitate big data analysts to formulate a real-time, intraoperative navigation system something like a surgical GPS [14].
- (D) Triad for AI-assisted surgeries – Three main contributions of a surgeon toward development of AI-powered ORs/OTs:
 1. Provide access to comprehensive data – The operating room is equipped with multiple data devices right from diagnostic radiographs; the Boyle's apparatus; the ECG and vital statistics monitoring device; the cardiac monitor/electrocautery unit; the pulse oximeter; the blood pressure, blood glucose, and body temperature recording devices; the suction irrigation; the operating light; the ultrasound; and many more devices. All these devices are constantly recording patient data and can

- serve as sources of input data or sensors of data. It is up to the surgeon to acclimatize the AI team and aid in feeding all the sensor data into the AI system so as to develop a real-time context-aware AI tool.
2. Explain or interpret the surgical data for the AI system to learn and improve its performance. The raw initial data needs constant annotation by surgeons. These annotations can be in the form of classifying data (e.g., which type of tissue is visible in the images, blood, bone, fascia, soft vs hard tissue, skin vs fascia, blood vessels by diameter, nerves vs blood vessels, etc.), semantic segmentation (e.g., linking pixel resolution to organs in an image), and linear numeric regression (e.g., defining objects by size) [15]. Data annotation does throw up challenges to the surgical and AI teams. These would be dealt with in the next segment (i.e., “challenges in the path”).
 3. Work in tandem with the machine learning systems – As the AI system begins to understand the data, it shifts into deep learning. This means the surgeon supervises the learning so that the system can now retain and apply information used to perform a former task (e.g., identifying a bleeder in the surgical site) to execute a succeeding related task (e.g., forewarning of risk and target structures via tissue navigation to prevent bleeders).

Challenges in the Path

1. *Annotation Challenges* – As discussed in the previous section, the AI system requires constant annotations from the surgeon:

- Time constraint is the foremost challenge as surgeons often have packed work schedules that demand their full attention.
- Furthermore, the surgical data is vast, and annotation demands expert knowledge, thus making it an expensive and time-consuming activity.

- A greater challenge of patient confidentiality and privacy arises as data from across many hospitals will be required to provide sufficient information to be representative of the task to be learned by the AI system.
- Moreover, there is absence of any standardized data acquisition and annotation policies as of now and minimal data in digital formats.

Suggestions for Overcoming This Challenge

- Active Learning: Annotation with selectively significant data points from among vast data sets [16].
 - Crowdsourcing: A method of problem-solving wherein a task is delegated to a team of individuals rather than a single or few people. The large team may contain relative novices, but their lack of experience is compensated by the distributed wisdom of the group members. Thereby, this approach is advantageous in terms of flexibility, efficiency, and scalability. Crowdsourcing has successfully been applied in healthcare in the diagnosis of colonic and polyps and identification of malaria-infected red blood cells and holds promise in the field of maxillofacial surgery too [17].
 - Ontologies: Help maintain consistency in the vocabulary used by surgeons for annotation, making it easily readable by the AI system.
2. *Policy Challenges* – Regulation and reimbursement schemes will determine the on-the-ground AI access and availability.

Suggestions for Overcoming This Challenge

There is not much a surgeon can directly do in this direction. However, policy-makers will definitely count the surgeon’s inputs as priority for formulating guidelines. Therefore, maintaining detailed standard operating procedure flow charts and notes on advantages of AI in surgery which are backed by evidence will prove beneficial.

3. Algorithm Complexity and Bias Challenges – AI will remain a novel subject to medical experts for a while. Implementation of AI systems will not be under the direct jurisdiction of medical personnel. Nor is the doctor/surgeon's primary duty to know the nuances and functioning of the AI system. Therefore, there exists the possibility of corporate or other authorities in power to camouflage wrong intents like discrimination by misuse of the complexity of the AI algorithmic functioning. An example of this scenario was a claim made against the auto company Uber by Tim Hwang and Madeleine Clare Elish of non-profit organization Data & Society. The allegation is that Uber uses surge pricing algorithms as pretense and commits unfair pricing by depicting a fake appearance of demand to the end user [18].

Suggestions for Overcoming This Challenge

- Appoint a board or committee of people from different specialties so as to avoid vested interests and abuse of possible algorithmic bias.
- Be informed of the basic principles of AI so as to be vigilant against malpractice.
- Ensure strict legal blueprint and government laws to root out unethical leverage of AI for malicious profit.

Literature Speaks: An Overview of Current Applications with Potential for Harnessing AI in Maxillofacial Surgery

Maxillofacial Presurgical Imaging

Current

- AI is being used both in image acquisition and interpretation to enhance diagnostic, prognostic, and risk analysis in surgery. Digital radiography has already taken over conventional methods, and incorporation of AI into intraoral scanning devices has brought about an added and much needed direct advantage apart from AI data acquisition – that of reduced patient

radiation exposure. AI ensures high-resolution images with lower radiation doses by optimizing signal-to-noise ratio [19].

- Quantification of tumor phenotype and decision-making support with the aid of “radiomics,” a method of AI that can analyze vast amounts of radiographic images to then extract relevant quantitative features from them [20].

Potential

- Deep learning techniques can be employed to learn from a standard hierarchical representation of a specific image from several repeated regular examinations to then differentiate normal and abnormal radiographic findings.
- Convolutional neural networks along with deep learning have the potential to identify disease characteristic features from radiographic images to predict nature of lesion/disease entity.

Orthognathic Surgery

Current

- Automated lateral cephalometric analysis with integrated AI and machine learning. One might misunderstand this application as mere laziness on the part of the surgeon. However, the AI cephalometrics is beyond 2D analytics. A true biometric facial analysis of a 3D cone beam radiograph (CBCT) requires 100–200 craniometric points. AI has unmatchable ability to analyze so many parameters simultaneously and accurately [21, 22].
- Dynamic virtual setup software such as Cin Check and Insignia [23] are powered by machine learning and effectively replace the tedious tasks of plaster model mock surgery and lateral cephalometric analysis. They further enhance orthodontic-surgical team communication, patient education, and treatment planning by means of visual imagery of each treatment objective and outcome possibilities.

Potential

- To extend machine learning employed in cephalometric analysis software to the treatment of sleep apnea by the use of AI-based 3D

reconstruction and superimposition of multiple diagnostic modalities such as patient's intraoral images, digital photographs, and CBCTs.

- Development of ambitious AI systems with the integration of 3D CBCT images, planned digital mock surgery osteotomy cuts and robotic arms to make intraoperative precise osteotomies and check dynamic patient occlusion on the table, and thereby highly superior facial recontouring with minimal hospitalization and recovery time.

Implant Surgery

Current

- A pilot study of a robotic arm placing a zygomatic implant in a phantom skull showed higher accuracy than conventional surgical placement. The authors wish to augment their findings in the future by addition of tactile sensors and other AI tools [24].

Potential

- Bone mineral density and osteoporosis can be diagnosed by AI, and this is clinically relevant to implant surgeries. Panoramic radiographs have been used by AI models to distinguish between osteoporotic and healthy subjects. The AI algorithm can evaluate alveolar bone cortical width and erosion of the cortex to distinguish between osteoporotic and healthy subjects with 95% accuracy, sensitivity, and specificity. These initial studies have been in favor of application of such AI models into routine surgical planning in the time to come [25].

Temporomandibular Joint (TMJ) Surgery

Current

- Japanese researchers have developed a chewing robotic skull that functions with wires and motors. It employs artificial muscle

actuators (AMAs) to simulate dynamic masticatory muscle activity and chewing. However, this is a generalized system more applicable for training and research on TMJ surgeries and will be difficult to build to replicate every individual patient [26]. However, this certainly can form a basis for future research into musculoskeletal disorders by making customized robotic skulls of individual patients with measurements from their CT scans and facial photographs/impressions of the head and neck region. This can be particularly useful for management of conditions such as Myofascial Pain Dysfunction Syndrome (MPDS), trigger point muscle pains, correction of abnormal jaw movements induced by muscle laxity or hypo/hyper activity. These conditions are often challenging to diagnose because of their multifactorial etiology, association with dental occlusion and inter-relation with TMJ functioning in the head and neck region.

- Internal derangements of the TMJs have been enigmatic and complex dilemmas due to several interrelated factors like occlusion, visualization of intra-articular space, interrelationship of both the TMJs (right and left), etc. Clinicians have had to rely on external clinical evaluation modes, thus making it difficult to arrive at a definitive diagnosis and optimal treatment. Turkish researchers have used data such as patient clinical symptoms and subsequent diagnosis of TMJ internal derangements to train artificial neural networks. The results have been promising with improved diagnostic efficacy [27].

Potential

- The surgical anatomy and approach to the TMJ is tricky owing to the proximity to significant anatomic structures like the parotid gland, the facial nerve, and branches of the external carotid artery. Therefore, residents and novice surgeons find it challenging to perform surgery in this region. AI and deep learning tools can come to the rescue of surgeons as navigation, facial plane recognition, and nerve sensor devices.

- Total alloplastic joint replacements are an emerging option in TMJ ankylosis management. AI can be leveraged in the precise fabrication and placement of these joints to improve outcomes especially related to range of movement of the alloplastic joint which currently is limited to only rotational movement.

Oncosurgery and Reconstruction

Current

- IBM Watson is an AI-based cognitive system that leverages natural language processing and dynamic learning to offer clinicians' evidence-based treatment options after analyzing patient reports. IBM Watson for oncology has been trained by experts at Memorial Sloan Kettering Cancer Institute and is currently in use in hospitals around the globe.
- Reconstruction of jawbones after oncosurgical resection is challenging keeping in mind esthetics and functionality of the head and neck region. Determining normal predisease jaw morphology after radical tumor surgery has been attempted with promising results by CTGAN a deep convolutional generative adversarial network (DCGAN) [28].

Potential

- It is a known fact that oncological surgeries are extensive time-taking procedures requiring patients to be under general anesthesia for several hours. AI systems for effective intraoperative management of patient vitals like blood pressure will provide immense relief to the surgical-anesthetic teams as seen in early-stage randomized controlled trials [29].
- Intelligent surgeon's scalpel/knife that integrated rapid analysis by AI systems to distinguish between tumor margins, healthy tissue, borderline, and malignant tissue will greatly improve the quality of oncosurgery. The amount of tissue removal will be optimized, intraoperative frozen section and time delay can be minimized, and superior closure and reconstruction can be achieved.

Trauma Surgery

Current

- Machine learning algorithms have predicted the postsurgical short-term outcomes of open reduction internal fixation (ORIF) in ankle fracture fixation surgeries [30].

Potential

- The incidence of occult fractures in maxillofacial traumatic injuries is high and a matter of concern to treating surgeons. Missed fractures and associated fractures that are difficult to detect on conventional radiographs often haunt the clinician. Machine learning systems can be beneficial in detection and isolation of occult fractures [31].
- AI deep learning guided selection of the most suitable internal fixation from among various hardware options for a given facial trauma and personalized to each patient by analysis of patient radiographs, age, and medical history, correlated to medical literature, for example, the type of plate to be used, the number of screws, plate versus lag screw, and one-point versus two-/three-point fixation. This predictive intelligence will prevent complications like malunion and asymmetry due to inadequate or overzealous fixation.
- AI computer vision is particularly helpful in panfacial trauma (multiple facial fractures) and ballistic (high-energy avulsive trauma, blast-induced) facial injuries. The challenge of sequencing in panfacial trauma and diagnosis of occult trauma and extent of tissue (soft and hard) loss can be greatly enhanced with computer vision navigation surgical systems.
- Comprehensive orbital and ocular trauma AI tool would be of great help in synchronizing patient care between ophthalmological and maxillofacial teams, thereby enabling swift care. Orbital fractures usually occur in conjunction with naso-ethmoidal and/or zygomatico-maxillary complex. Furthermore, thorough ophthalmological evaluation in terms of visual acuity, visual fields, fundoscopic examination, papillary responses, and ocular motility is

necessary prior to surgical reduction of these fractures. An AI tool can be trained to accurately perform optic tests and be of great help to novice surgeons or trauma teams with limited access to ophthalmic devices.

Impacted Teeth and Minor Oral Surgery

Current

- Disimpaction surgeries are the most common procedure performed by maxillofacial surgeons. Injury to critical anatomic structures like the inferior alveolar nerve during impacted wisdom teeth removal has been a long-standing and highly studied aspect. Other minor oral surgeries like alveo-loplasties, tori removal, salivary mucocele excisions, etc. could also be improved with novel visualization and guided surgical systems. An augmented reality tool was developed and tested to visualize the inferior alveolar nerve (IAN) bundle through the mandibular bone by the use of a fiducial marker on an occlusal splint. A virtual image of the surgical site is created by 3D software analysis of patient's CT images.

An integrated image is then achieved by superimposition of the virtual image on the real surgical environment via the occlusal splint-fiducial marker assembly. The results were promising in enabling protection to the IAN and superior surgical outcomes [32].

- Prediction of postoperative facial swelling and edema which psychologically and physically affect a patient's recovery after minor oral surgery by artificial neural networks [33] has been far superior to traditional methods of prediction.

Potential

- Deep learning solutions to a commonly encountered challenge of management of medically compromised patients in oral surgery. The use of AI to evaluate relevant and current literature on precautions and protocols for management of patients on anticoagulants, antihypertensives, thyroid medication, and

other pharmacological agents will be highly assistive to maxillofacial surgery.

Miscellaneous

Facial Paralysis Gradation – Facial nerve paralysis due to trauma, infection, iatrogenic causes, and complications of surgery in parotid/TMJ regions resulting in unilateral facial palsy-/Bell's palsy-like condition. Management of such cases hinges on effective grading of the extent of neuronal injury. Currently there exists no objective dependable classification of facial nerve paralysis. The use of video footage of patients performing assigned facial movements to test nerve function can serve as data for artificial neural networks to effectively ascertain the degree/grade of facial nerve paralysis [34].

Pain Control – Maxillofacial surgery is particularly a bridge specialty of dental and medical fields, and its procedures routinely involve both local and general anesthesia and conscious sedation and ambulatory anesthesia. AI for optimal anesthesia delivery by automating drug delivery based on machine learning predictions of drug pharmacokinetics within the body has been reported. Neural networks have been used to predict recovery rates from anesthesia. Machine learning techniques are being tested to monitor conscious sedation and alert about levels of respiratory depression and automate classification of patients' preoperative ASA (acuity) status [35].

Watch List for Future Forward Maxillofacial Surgeons

Surgical Data Science

Large amount of annotated data for training the AI system is required. This has led to the emergence of a unique field of surgical data science. This involves the use of AI tools to capture, organize, and analyze which in turn will fuel the AI system to provide AI-enhanced operative room experience. Such AI systems would improve the quality

of surgical care by providing real-time decision support, context-aware assistance, and cognitive robotics [36].

Surgical Scene Analytics

Recent studies were focused on AI analysis of surgical tools to enable automated surgical (robotic) instrument usage. However, the future will see expansion of this methodology into entire surgical scene analytics. This could translate into anything from a virtual second assistant to a live intraoperative prediction device [37].

Surgical Control Tower (SCT)

Artificial intelligence developed by studying human detection poses estimation and surgical procedure monitoring via wall-/ceiling-mounted camera image data. The SCT will serve as centralized control for all round assistance to the surgical team. The ongoing OR status can be analyzed to provide smart decisions and automation of tasks and reduce errors in complex situations. The SCT can also serve as a video reservoir of data for training and evaluation [38].

Conclusion

The evolution of maxillofacial surgery has been shaped by time from the first accounts of mandibular fracture managements in the 2700 BC Egyptian *Edwin Smith Papyrus* [39] to the state-of-the-art “Fibula Jaw in a Day” maxillofacial reconstructions [40]. Fortunately the fraternity that emerged from medicine and merged with dentistry into a unique bridging specialty has been constantly upgrading and adapting to enhance practice. Training paradigms are now shifting from conventional modalities, and the most dominant influence has been that of technology. The scope of AI in maxillofacial surgery is poised to witness practical applications in every aspect of the field – right from AI-powered automation of dental extractions and robotic AI-enhanced surgical instrumentation to

machine learning-enabled decision support in diagnosis of enigmatic and challenging medical conditions such as genetic syndromes of the head and neck. The applications are infinite, and this chapter has delved into the current literature and guidelines for future potentials of AI integrated maxillofacial surgery.

References

1. Poole DL, Mackworth A, Goebel RG. Computational intelligence and knowledge. In: Computational intelligence: a logical approach. Oxford University Press; 1998. p. 1–22.
2. Miller DD, Brown EW. Artificial intelligence in medical practice: the question to the answer? *Am J Med.* 2018;131(2):129–33.
3. Buscema M. A brief overview and introduction to artificial neural networks. *Subst Use Misuse.* 2002;37:1093–148.
4. Fasel B. An introduction to bio-inspired artificial neural network architectures. *Acta Neurol Belg.* 2003;103:6–12.
5. Lisboa PJ, Taktak AF. The use of artificial neural networks in decision support in cancer: a systematic review. *Neural Netw.* 2006;19:408–15.
6. Markuzon N, Carpenter GA. ARTMAP-IC and medical diagnosis: instance counting and inconsistent cases. *Neural Netw.* 1998;11:323–36.
7. Payne SJ, Arrol HP, Hunt SV, Young SP. Automated classification and analysis of the calcium response of single T lymphocytes using a neural network approach. *IEEE Trans Neural Netw.* 2005;16:949–58.
8. Subasi A, Alkan A, Koklukaya E, Kiyimik MK. Wavelet neural network classification of EEG signals by using AR model with MLE preprocessing. *Neural Netw.* 2005;18:985–97.
9. Liu D, Xiong X, Hou ZG, Dasgupta B. Identification of motifs with insertions and deletions in protein sequences using self-organizing neural networks. *Neural Netw.* 2005;18:835–42.
10. Yang ZR, Thomson R. Bio-basis function neural network for prediction of protease cleavage sites in proteins. *IEEE Trans Neural Netw.* 2005;16:263–74.
11. Rosandich RG. HAVNET: a new neural network architecture for pattern recognition. *Neural Netw.* 1997;10: 139–51.
12. Gao Y, Er MJ. An intelligent adaptive control scheme for postsurgical blood pressure regulation. *IEEE Trans Neural Netw.* 2005;16:475–83.
13. Natarajan P, Frenzel JC, Smaltz DH. Demystifying big data and machine learning for healthcare. Boca Raton: CRC Press; 2017.
14. Pereira KR, Sinha R. Welcome the “new kid on the block” into the family: artificial intelligence in oral and maxillofacial surgery. *Br J Oral Maxillofac Surg.* 2020;58:83–4.

15. Bodenstedt S, Wagner M, Müller-Stich BP, Weitz J, Speidel S. Artificial intelligence-assisted surgery: potential and challenges. *Visc Med.* 2020;36:450–5.
16. Breucha M, Müller-Stich B, et al. Active learning using deep Bayesian networks for surgical workflow analysis. *Int J CARS.* 2019;14(6):1079–87.
17. Dai JC, Lendvay TS, Sorensen MD. Crowdsourcing in surgical skills acquisition: a developing Technology in Surgical Education. *J Grad Med Educ.* 2017;9(6):697–705.
18. Tim Hwang and Madeleine Clare Elish, “The Mirage of the Marketplace,” Slate, August 9, 2015. http://www.slate.com/articles/technology/future_tense/2015/07/uber_s_algorithm_and_the_mirage_of_the_marketplace.html
19. Sun Y, Liu X, Cong P, Li L, Zhao Z. Digital radiography image denoising using a generative adversarial network. *J Xray Sci Technol.* 2018;26(4):523–34. <https://doi.org/10.3233/XST-17356>.
20. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology.* 2016;278:563–77.
21. Nishimoto S, Sotsuka Y, Kawai K, Ishise H, Kakibuchi M. Personal computerbased cephalometric landmark detection with deep learning using cephalograms on the Internet. *J Craniofac Surg.* 2019;30(1):91–5. <https://doi.org/10.1097/SCS.0000000000004901>.
22. Zamora N, Llamas J-M, Cibrián R, Gandia J-L, Paredes V. A study on the reproducibility of cephalometric landmarks when undertaking a three-dimensional (3D) cephalometric analysis. *Med Oral Patol Oral Cir Bucal.* 2012;17(4):e678–88. <http://www.ncbi.nlm.nih.gov/pubmed/22322503>
23. Hennessy J, Al-Awadhi EA. Clear aligners generations and orthodontic tooth movement. *J Orthod.* 2016;43(1):68–76. <https://doi.org/10.1179/1465313315Y.0000000004>.
24. Zhenggang C, Qin C, Fan S, Yu D, Wu Y, Chen X. Pilot study of a surgical robot system for zygomatic implant placement. *Med Eng Phys.* 2019;75:72.
25. Vlasiadis KZ, Damilakis J, Velegrakis GA, Skouteris CA, Fragouli I, Goumenou A, et al. Relationship between BMD, dental panoramic radiographic findings and biochemical markers of bone turnover in diagnosis of osteoporosis. *Maturitas.* 2008;59:226–33. <https://doi.org/10.1016/j.maturitas.2008.01.006>.
26. Takanishi A, Tanase T, Kumei M, Kato I. Development of 3 DOF jaw robot WJ-2 as a human’s mastication simulator. In: Proceedings of the international conference on advanced robotics ICAR, p. 277–82, Pisa. 1991.
27. Bas B, Ozgonenel O, et al. Use of artificial neural network in differentiation of subgroups of temporomandibular internal derangements: a preliminary study. *J Oral Maxillofac Surg.* 2012;70(1):51–9.
28. Liang Y, Huan J, Li JD, et al. Use of artificial intelligence to recover mandibular morphology after disease. *Sci Rep.* 2020;10:16431.
29. Wijnberge M, Geerts BF, Hol L, Lemmers N, Mulder MP, Berge P, et al. Effect of a machine learning-derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the HYPE randomized clinical trial. *JAMA.* 2020;323(11):1052–60.
30. Merrill RK, Ferrandino RM, Hoffman R, Shaffer GW, Ndu A. Machine learning accurately predicts short-term outcomes following open reduction and internal fixation of ankle fractures. *J Foot Ankle Surg.* 2019;58:410.
31. Hendrickx LAM, Sobol GL, Langerhuizen DWG, Bulstra AEJ, Hreha J, Sprague S, Sirkin MS, Ring D, Kerkhoffs GMMJ, Jaarsma RL, Doornberg JN. Machine learning consortium. A machine learning algorithm to predict the probability of (Occult) posterior malleolar fractures associated with tibial shaft fractures to guide “Malleolus first” fixation. *J Orthop Trauma.* 2020;34(3):131–138.
32. Zhu M, Liu F, Chai G, et al. A novel augmented reality system for displaying inferior alveolar nerve bundles in maxillofacial surgery. *Sci Rep.* 2017;7:42365.
33. Zhang W, Li J, Li ZB, Li Z. Predicting postoperative facial swelling following impacted mandibular third molars extraction by using artificial neural networks evaluation. *Sci Rep.* 2018;8(1):12281.
34. McGrenary S, O'Reilly BF, Soraghan JJ. Objective grading of facial paralysis using artificial intelligence analysis of video data. In: 18th IEEE symposium on computer-based medical systems (CBMS'05), Dublin, 2005, p. 587–92.
35. Hashimoto DA, Witkowski E, Gao L, Meireles O, Rosman G. Artificial intelligence in anesthesiology: current techniques, clinical applications, and limitations. *Anesthesiology.* 2020;132:379–94.
36. Maier-Hein L, Vedula SS, Speidel S, Navab N, Kikinis R, Park A, et al. Surgical data science for next-generation interventions. *Nat Biomed Eng.* 2017;1(9):691–6.
37. Allan M, Kondo S, Bodenstedt S, Leger S, Kadkhodamohammadi R, Luengo I, et al. Robotic scene segmentation challenge. 2020. arXiv:2001.11190.
38. Padoy N. Machine and deep learning for workflow recognition during surgery. *Minim Invasive Ther Allied Technol.* 2019;28:82–90.
39. Laskin DM. Oral and maxillofacial surgery: the mystery behind the history. *J Oral Maxillofac Surg Med Pathol.* 2016;28(2):101–4.
40. Qaisi M, Kolodney H, Swedenburg G, Chandran R, Calossi R. Fibula jaw in a day: state of the art in maxillofacial reconstruction. *J Oral Maxillofac Surg.* 2016;74(6):1284.e1–1284.e15.
41. Massilla Mani F, Sivasubramanian SS. A study of temporomandibular joint osteoarthritis using computed tomographic imaging. *Biom J.* 2016;39(3):201–6.



Mauricio do Nascimento Gerhardt, Sohaib Shujaat, and
Reinhilde Jacobs

Contents

Introduction	906
AI for Assessment in Dentistry.....	907
AI for Diagnosis in Dentistry.....	911
AI for Treatment Planning	912
AI for Outcome Prediction in Dentistry.....	912
Concluding Remarks	914
References	916

Abstract

The steep rise of digital dentistry and technological advancements have opened doors for the development of artificial intelligence (AI). For the past few years, AI-based applications in dentistry have been constantly evolving as highlighted by the increasing number of studies, and now it is slowly entering the clinical arena. As healthcare professionals, dentists need to diagnose, plan, and make clinical decisions in order to provide an adequate treatment and care for their patients. All these phases are time-consuming, observer-dependent, and subjected to human error. Currently, the studies applying AI in many dental specialties have validated its application for the purpose of diagnosis and clinical decision-making. Thus, the objective of AI is to combine the professional expertise with the computer-assisted systems to automatize complex tasks, mimic

M. do Nascimento Gerhardt

OMFS IMPATH Research Group, Department of Imaging and Pathology, University of Leuven and Oral & Maxillofacial Surgery, University Hospitals Leuven, KU Leuven, Leuven, Belgium

School of Health Sciences, Faculty of Dentistry, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, Brazil

S. Shujaat

OMFS IMPATH Research Group, Department of Imaging and Pathology, University of Leuven and Oral & Maxillofacial Surgery, University Hospitals Leuven, KU Leuven, Leuven, Belgium

e-mail: sohaib.shujaat941@gmail.com

R. Jacobs (✉)

OMFS IMPATH Research Group, Department of Imaging and Pathology, University of Leuven and Oral & Maxillofacial Surgery, University Hospitals Leuven, KU Leuven, Leuven, Belgium

Department of Dental Medicine, Karolinska Institute, Stockholm, Sweden

e-mail: reinhilde.jacobs@kuleuven.be

human cognitive skills, and retrieve information from digital data. Dental AI applications can be advantageous for all dental specialties including dentomaxillofacial radiology, restorative dentistry, oral and maxillofacial surgery, orthodontics, periodontics, prosthodontics, endodontics, and forensic dentistry.

Even though most researches and developments are still in an early phase, current results in the dental field are encouraging for future clinical applications.

This chapter provides an overview of the current state of the art of the AI applications in dentistry and its specialties.

Keywords

Digital dentistry · Presurgical planning · Cone-beam computed tomography · Intraoral scanner · Panoramic radiography · Radiological diagnosis · Tooth · Jaw · Face

Introduction

The daily routine tasks of a dentist involve gathering patient information, diagnostics, and planning and performing clinical treatment procedures. At the first visit to a dental office, the dentist needs to collect and annotate many pieces of information in a detailed manner. Afterward, the clinician initiates the diagnostic phase, combining clinical and imaging data, to arrange all information in a dental chart. All these tasks are time-consuming and completely dependent on the professional's attention to possible minor details. Furthermore, full integration of these multiple sources of information and data is crucial. Only after careful collection and integration of all information and diagnostic data, the dentist will be able to start further clinical treatment procedures.

During the past decade, technological advancements have already provided some solutions such as electronic medical and dental records, software applications to plan oral rehabilitation, and augmented reality to visualize the outcomes of the planned treatment. However, these solutions still depend on human intervention

and demand a lot of time. In this context, it would be important to develop more efficient software to allow reliable and automated actions for aiding the dentists in their assessments, treatment planning, and decision-making strategies. This is where artificial intelligence (AI) can jump in to facilitate digitalization and automation.

AI is rapidly evolving in dentistry with the number of studies and applications drastically increasing worldwide during the past few years. The storage of data and information in dentistry has always been a part of the daily clinical routine, as dentists are often collecting imaging data accompanied by general and dental history, systemic conditions, and medications in their patient's records. Nowadays, it has become much easier with digital data storage. This makes dentistry a suitable field for the implementation of new technologies such as AI and its subclasses, machine learning (ML) and deep learning (DL). The aid of technology may play a substantial role in terms of assessment, diagnosis, planning, clinical decision-making, and outcome prediction, consequently improving prognosis as well. Some studies indicate that the AI-based systems are able to surpass even the performance of dental specialists [1, 2].

In dentistry, facial and dental aesthetics are of a major concern. With the improvement of dental materials and techniques, patient's expectations in relation to the outcomes of the treatment have drastically increased. Therefore, accurate diagnosis and treatment planning are essential when dealing with tissues and anatomical structures that have an impact on people's appearance and also to provide personalized healthcare taking into account the patient's unique characteristics.

Thus, we must comprehend that AI is an advancing science that does not intend to replace dental specialists but to act as a powerful ally for attaining a better, faster, and more accurate assessment and diagnosis of the cases with the empathy and care given by the dentists. Finally, it intends to provide better outcomes for patients at lower costs. This combination of technology and dental practitioner is called augmented intelligence [2, 3].

The aim of this chapter is to give the reader an overview of the state-of-the-art AI applications in

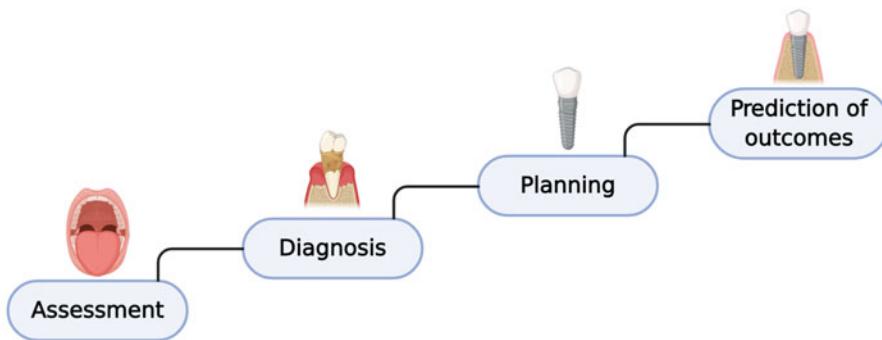


Fig. 1 Steps of the dental treatment where AI can aid dental specialists

dentistry, following a progressive timeline of dental treatments, in terms of assessment, diagnosis, planning, and prediction of outcomes in dentistry (Fig. 1).

AI for Assessment in Dentistry

Currently, one of the fields that has gained most attention for the use of computer systems aiming at improving tasks performed in dentistry is oral radiology. As the computer systems are able to extract and quantify features which the human eye cannot detect, thereby turning radiology into an objective science. In this context, AI and its subclasses have been excellent tools for serving radiologists to analyze the amount of information contained in different types of images. DL has been applied for many purposes such as optimization of images with low radiation dose and less scattering or artifacts, classification, image registration, segmentation, lesion detection, image retrieval, and image-guided therapy. It can also be applied for correcting technical positioning errors on panoramic radiographs and developing patient-specific imaging protocols [3].

Imaging-based data are also ideal for further development of AI-based applications, mainly for treatment, surgical planning, and follow-up. One of the first steps in order to train an automated tool is the segmentation of anatomical structures. This is usually a manual, time-consuming, and operator-dependent task, which is subjected to human error. Automation of this work may

improve diagnosis and treatment planning, virtual surgery planning, radiation therapy planning, and radiomics analysis [4, 5].

AI has been used for identification of teeth using different radiological imaging modalities. A variety of studies reported on the use of automated systems to effectively recognize and label teeth on periapical, bite-wing, panoramic, and cone-beam computed tomography (CBCT), and they achieved results comparable to those obtained by specialists [6–9]. Zhang et al [6], and Chen et al [7], applied DL algorithms for recognizing and labeling teeth position on intraoral radiographs and achieved high levels of precision. The former had 95.8% and the latter exceeded 90% precision. Tuzoff et al [8], also applied DL for teeth detection and numbering on panoramic radiographs and achieved a sensitivity of 0.99 and precision of 0.99 for detection and a sensitivity of 0.98 and specificity of 0.99 for numbering. Leite et al [9], used an AI-driven tool for tooth detection and segmentation on panoramic radiographs, and the system achieved 98.9% sensitivity and 99.6% precision for tooth detection. They also highlighted that the method was significantly faster, consuming 67% less than the time needed for manual processing. A panoramic radiograph is commonly requested for case evaluations because of the general overview provided. An example of an AI-driven tool for tooth detection and segmentation is presented in Fig. 2. Similarly, the application of AI for tooth recognition and segmentations has also been explored three-dimensionally using CT and CBCT devices

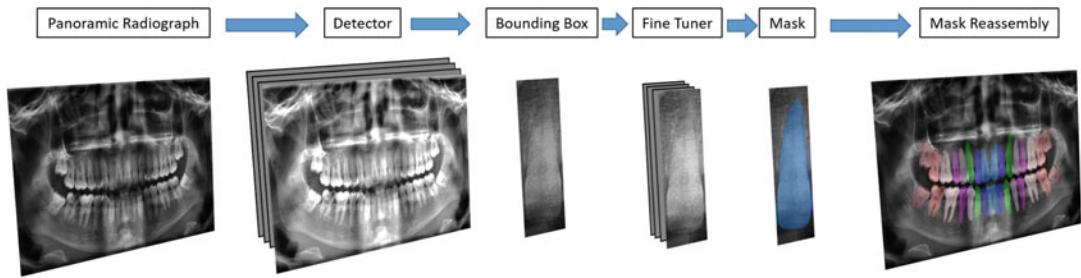


Fig. 2 Workflow of an AI-driven tool for tooth detection and segmentation [4]

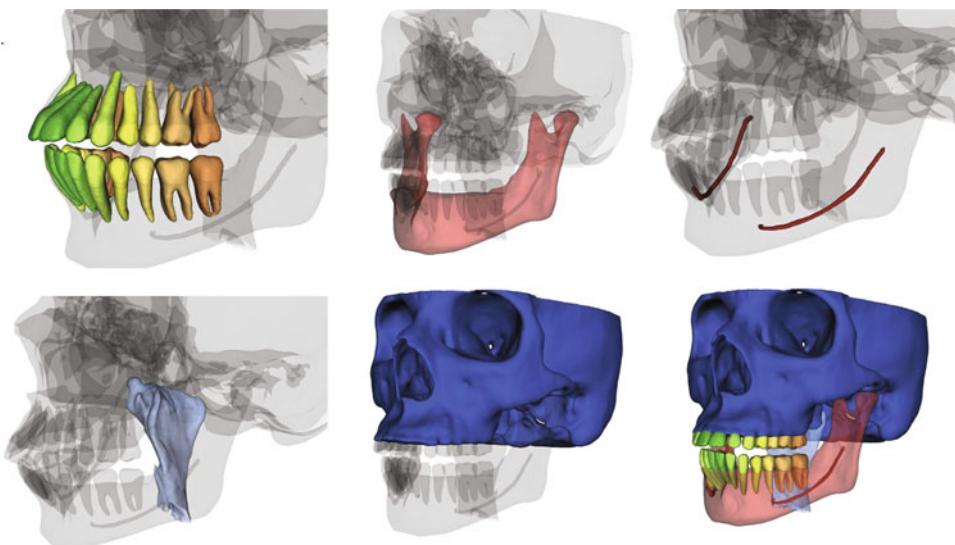


Fig. 3 Dentomaxillofacial structures automatically identified and labeled on a CBCT image (Relu BV, Leuven, Belgium)

[10–12]. Figure 3 illustrates an example of AI-based segmentation of dentomaxillofacial structures. All these results suggest that AI could be efficiently applied to recognize and label teeth on dental charts in a time-saving manner.

In the field of implant dentistry, the AI-based identification of dental implants may be another important task as they have been the treatment of choice for oral rehabilitation for more than 30 years. Since then, many different brands and systems have been developed. With the growing number of implants being inserted globally, the reports of complications have also increased. In order to solve a problem with a dental fixture, it is necessary to recognize the implant manufacturer and its prosthetic connection system. In some cases, the identification of implant is difficult

because the information may be lacking in the dental records or the patient is attending another professional that does not know the implant by its characteristics. Therefore, AI may be a useful tool to identify dental implants from radiographic exam [13–15]. Takahashi et al [13] applied DL in a pilot study to identify implants among six different systems manufactured by three companies. The results suggested that dental implants can be identified from panoramic radiographs. Hadj Saïd et al. [14] also applied DL to identify six types of implants from three manufacturers and achieved almost 94% diagnostic accuracy. Lee and Jeong [15] applied DL to recognize three similar conical dental implants from panoramic and periapical radiographs and compared the results with the performance of a periodontist.

Results showed that DL achieved an area under the curve (AUC) of 0.97 (95% CI, 0.96–0.98), while the professional values were 0.92 (95% CI, 0.91–0.93). The sensitivity and specificity were 95% and 97%, respectively, for the former and 88% and 87% for the latter. Even though the results of both studies were encouraging, they either relied on a small dataset or needed a data augmentation procedure. Therefore, if we need to apply these types of AI tools in a routine clinical practice, then their performance should be validated with all implant systems using multiple 2D and 3D imaging modalities and scanning parameters.

In relation to the implant placement in the lower jaw, a surgeon must be aware of the mandibular canal (MC) localization for determining the implant dimensions without damaging the nerve. The identification and segmentation of MC are normally carried out manually by the radiologists on 3D CBCT images. Jaskari et al [4] demonstrated that the use of DL to segment MCs significantly reduced the radiologists' time consumption for the task compared to manual delineation. Results in terms of accuracy were within 0.5 mm for about 90% of MCs. Toward the future, such AI-driven neurovascular canal detection and segmentation need to be trained with more anatomical variabilities and images derived from different CBCT units.

One of the most common conditions encountered in a routine dental practice is the impacted third molar, and the dentists are required to predict

whether the tooth will erupt in the oral cavity normally or would require extraction. The two important factors for predicting the potential eruption of these teeth are their angulation and the space between the distal side of the second molar and the anterior border of the ramus. These measurements can be performed on a panoramic radiograph by a radiologist. However, like every other task mentioned before, it is a tedious and time-consuming task. Vranckx et al [16] applied an AI-based algorithm for automatically segmenting and determining the angulation of the lower molars on panoramic radiographs. They measured the performance of the AI tool and the time required for the task versus an expert. Their results indicated that the AI tool had a good performance (Intersection over Union, 0.88) for segmentation and angulation calculation. Additionally, it was found to be twice as fast as the manual measurements. An example of the automated segmentation and angulation measurement of the lower molars on a panoramic radiograph is shown in Fig. 4.

When it comes to orthodontic treatment, lateral cephalometric analysis is routinely performed, and it consists of the identification of skeletal and soft tissue landmarks with the objective to assess various angles, distances, and ratios and to plan and predict the treatment outcomes. The tracing of these landmarks is a tedious activity, and manual identification still remains the gold standard. This task is observer-dependent and

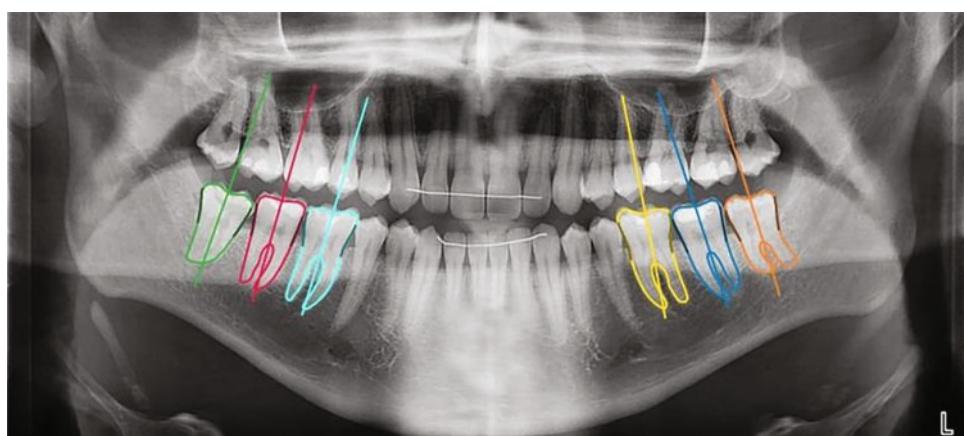


Fig. 4 Automated tooth angulation measurement on a panoramic radiograph [4]

varies depending on the experience level of an orthodontist. Inaccurate identification of the landmarks may lead to inappropriate decision-making for orthodontic therapy. It has been demonstrated that AI and DL algorithms have the potential to recognize anatomical landmarks and automatically detect the skeletal classification on cephalometric radiographs accurately [17–20]. During assessment of an orthodontic case, it is also crucial to determine the chronological age and to estimate the skeletal maturation of the patient. Both these factors are important for evaluating the remaining growth and development potential which can influence the orthodontic plan. Some studies have investigated the use of AI to perform these tasks [21]. All the aforementioned tasks such as radiographic analysis and identifying anatomical landmarks are time-consuming tasks for both clinicians and radiologists. Thus, the automation of these tasks would aid clinicians, as no manual intervention would be required when filling dental charts or elaborating radiographic reports.

In periodontology, periodontitis is one of the most commonly observed diseases in the oral cavity which can lead to alveolar bone loss and edentulism. This in turn can influence the patient's quality of life and impose higher healthcare costs, thereby requiring an early and accurate diagnosis. Evidence suggests only a few studies applying AI for the diagnosis of periodontal diseases. Krois et al. [22] and Kim et al. [23] proposed a convolutional neural network (CNN)-based network for detecting periodontal bone loss on panoramic radiographs. Both studies showed variable findings, where Krois et al [22] found higher accuracy in favor of the CNN (0.67) compared to the examiners (0.76) without any significant difference, whereas Kim et al. [23] suggested higher accuracy with CNN (0.75), and the average score of the dental clinicians was found to be 0.69. In another study, Lee et al. [24] developed a CNN algorithm for diagnosing and predicting bone loss on periapical radiographic images. They found the network to be as effective as an experienced periodontist. Recently, in another study,

Chang [25] et al. applied CNN for diagnosing as well as staging periodontitis based on panoramic radiographs. They demonstrated a high accuracy and stronger correlation between the CNN and an experienced radiologist compared to inexperienced ones, thereby confirming its effectiveness and reliability. As the 2D radiographs offer different levels of magnification and distortion, hence, there is a need to optimize the images further for training the CNN algorithms before employing these systems in a daily practice.

In endodontics, localization of the apical foramen is essential for determining the working length (WL) of the pulp canals accurately. Thereby, an inaccurate WL might either lead to an incomplete removal of the pulp or invasion of the periapical tissues. Saghiri et al [26]. conducted an ex vivo study and applied artificial neural network (ANN) for evaluating the position of a file tip in relation to the minor apical foramen and determining the WL on radiographs. They compared the AI performance with professional assessment and found that the AI-based network correctly determined the WL in 96% of the cases compared to the 76% achieved by the endodontists. Based on these findings, AI proved to be an excellent adjunct tool to locate the apical foramen and determine the WL.

A relatively new field where AI can play an important role is forensic dentistry. Sometimes, identifying victims of great disasters, such as plane crashes, is only possible with the analysis of the dental remains, and a forensicist has to spend a huge amount of time to identify and analyze the data. Similarly, another common task in forensic dentistry is the estimation of adolescent or young adult age by evaluation of the third molar development stage; however, this assessment is susceptible to human error and observer variability. The automated identification of the dental remains, the estimation of the third molar stage, and the identification of morphometric parameters based on panoramic radiographs may help to distinguish a person's identity, age, and gender. Some AI-based early phase studies have been carried out in the aforementioned areas,

thereby confirming that in near future, AI might be considered useful in the resolution of disasters and age estimation [27, 28].

AI for Diagnosis in Dentistry

The most beneficial impact observed with AI in the dental diagnostics has been the improvement in detecting lesions which might get misdiagnosed or overlooked by radiologists or clinicians on dental radiographs, such as caries, apical lesions, Sjögren's syndrome, sinusitis, extranodal extension of cervical lymph node metastases, and osteoporosis [3].

Dental caries is a common chronic infectious disease affecting many people around the world. Although the restoration methods to save teeth have significantly improved over the past few years, the same has not happened with the diagnostic methods. If a tooth cavity is detected at an early stage, the resolution prognosis is more optimistic. However, in the presence of deep fissures, tight interproximal contacts, and secondary lesions under restorations, the detection may be difficult. Even though dental radiography and the use of a dental probe could be considered as excellent tools for detecting dental caries, nevertheless, the final diagnosis tends to be based on empirical evidence. AI-based methods have been applied to detect dental caries, which may represent a vital shift to an objective way for diagnosing this type of lesion. Devito et al [29]. applied an AI model to detect proximal caries on bite-wing radiographs. The results suggested that AI was able to increase the diagnostic accuracy by 39.4% when compared with the mean of the examiners. Lee et al [30]. also applied a DL-based CNN to detect dental caries on periapical radiographs, and the algorithm achieved a mean AUC of 0.89. These results show how beneficial the use of AI can be for detecting one of the most common lesions seen by a dentist. Furthermore, allowing a change in the subjective diagnostic perspective which is completely dependent on the professional's experience to an objective one.

The correct diagnosis of crack or root fracture is crucial for making an accurate treatment strategy which may either involve tooth extraction or a more conservative approach. This task is difficult even on CBCT images where the fracture might not be visible to a naked eye, resulting in misdiagnosis and an inaccurate treatment approach. Some studies investigated the application of AI to detect root fractures effectively [31, 32]. Johari et al [31]. applied a probabilistic neural network to detect vertical root fractures (VRFs) on intact and endodontically treated teeth using periapical and CBCT radiographs. The results showed a high accuracy of 96.6%, and the authors concluded that the designed neural network can be used to diagnose VRFs on intact and endodontically treated teeth. CBCT images were more effective than periapical radiographs when evaluating the VRFs. Fukuda et al [32]. also applied CNN to detect VRFs on panoramic radiographs. The AI model achieved precision of 0.93, and it was found to be more effective than the specialists.

Some oral diseases can be life-threatening if not detected or treated early. Among such diseases, oral and maxillofacial cancer sits on the top of the pyramid having a high mortality rate. One of the most common oral malignancies is oral squamous cell carcinoma (OSCC), which presents with a high incidence rate worldwide, and most of the cases are diagnosed at an advanced stage, thereby reducing the survival rate of the patients. Aubreville et al [33]. evaluated an automatic approach for diagnosing OSCC by applying a DL approach to confocal laser endomicroscopy (CLE) imaging. They compared this approach against a textural feature-based ML method and obtained an AUC of 0.96 and a mean accuracy of 88.3%. The authors highlighted that this noninvasive method has the advantage of high magnification and better depth penetration allowing the diagnosis of the malignant tissue 100 μ m below the surface. The automated real-time identification of potential malignant sites in the oral mucosa may help to speed up the treatment and improve the prognosis.

Some jaw tumors vary in relation to the invasiveness and aggressiveness; however, they show similar radiological characteristics. Thereby, treatment is based on the differential diagnosis rather than an absolute diagnosis which might have a negative impact on the treatment plan. For instance, keratocystic odontogenic tumors (KCOTs) do not require radical jaw segmentation but enucleation like other cystic lesions. On the other hand, ameloblastomas require extensive surgical removal and reconstruction. Appropriate diagnosis and radiological differentiation between these two tumors can guide oral and maxillofacial surgeons to plan the treatment precisely. Poedjiastoeti and Suebnukarn [34] applied CNN for classification of KCOTs and ameloblastomas on digital panoramic radiographs. The results showed 83% in accuracy, which was comparable to the specialist (82.9%) but, however, much faster (38 seconds against 23.1 minutes).

AI for Treatment Planning

AI has also been useful for planning in dentistry. For orthodontic treatment, AI can be applied to decide if orthodontic treatment is needed or when to extract a tooth or not prior to the installation of fixed appliance [35–37]. In addition, orthodontic treatment might require surgical intervention such as orthognathic surgery to correct skeletal abnormalities of the maxilla and mandible and has to be carefully and assertively planned. It has been demonstrated that AI can decide between surgery and non-surgery cases based on lateral cephalometric radiographs [38].

In case of a surgical treatment, the advancements in technology have enabled combining imaging data with virtual models. This procedure requires an accurate superimposition of the data through the recognition of common reference points. Although some intrinsic and extrinsic registration methods have been proposed, nevertheless, it is time-consuming and requires

specific knowledge of certain software programs to achieve such a task. Meanwhile, registration can be achieved with AI in a user-friendly and time-efficient manner, thereby allowing accurate measurements and fabrication of digital splints for orthognathic surgery, oral and maxillofacial surgery appliances, dental implant insertion guides, navigated surgery, and many other applications.

AI for Outcome Prediction in Dentistry

One of the most important predictive activities would be certainly to detect a disease, such as oral cancer, in an early phase before it metastasizes, providing the patient with an adequate treatment and an excellent prognosis. DL can be applied for this purpose as it learns features from the training data to cast predictions on unseen data.

Patients frequently ask their dentists about the outcomes of the treatments which is a difficult answer to provide, as the predictive outcomes are influenced by a number of factors that would demand a certain amount of data and statistical analysis before giving the patient a proper response. However, some studies have focused on the use of AI to predict the outcomes of treatments such as postoperative swelling following the extraction of the third molar [39].

For orthodontic cases, depending on the circumstances during the mixed dentition phase, the orthodontist needs to save or open space for unerupted teeth. A study by [40] investigated the use of ANN to predict the size of unerupted canines and premolars of children with mixed dentition. The ANN selected the mandibular first molar and incisors and the maxillary central incisors for predicting the width of canines and premolars in the mandible and maxilla, respectively. The prediction error rates and maximum rates of over-/underestimation were smaller than the linear regression [40].

Orthognathic surgery aims at correcting severe dentofacial discrepancies that cannot be fixed

with conventional orthodontic treatment. Such procedures change the facial aesthetics which is of a major concern for the patients. AI-based approach has been applied to predict facial soft tissue changes in patients undergoing orthognathic surgery. Lu et al [41]. applied ANN to improve the prediction of facial changes on profile video images of post-orthognathic surgery patients. The ANN enhanced the prediction ability by more than 80%. An accurate prediction of surgical outcomes might improve the treatment planning and ability to provide patients with an accurate description of the changes following surgery.

AI has been widely implemented in conservative dentistry, dental implantology, and prosthodontics for predicting the longevity of restorations and their potential debonding, prediction of ceramic recipe in order to match the color of

natural teeth and implant treatment success rate and soft tissue transformations after rehabilitation with complete dentures [42–46]. The capability of predicting outcomes is based on many factors such as [1] characteristics of the restoration needed, [2] patient's personal habits, [3] dentist's abilities, and [4] properties of the restorative materials. A prediction of all these factors with AI might improve the decision-making process. The way forward is to implement these predictive AI-driven networks at a broader level as it has the ability to streamline dental care by reducing the laborious routine tasks of the dentists, predicting outcomes, and allowing an accurate treatment plan based on the AI-based prediction model.

A summary of the artificial intelligence applications reported in this chapter is provided in Table 1.

Table 1 Summary of AI applications in dentistry

Study	Field	Application
Zhang et al. (2018) [6]	Oral and maxillofacial radiology	Recognition of teeth on intraoral periapical radiographs
Chen et al. (2019) [7]	Oral and maxillofacial radiology	Detection and numbering of teeth on intraoral periapical radiographs
Tuzoff et al. (2019) [8]	Oral and maxillofacial radiology	Automatic teeth detection and numbering on panoramic radiographs
Leite et al. (2020) [9]	Oral and maxillofacial radiology	Automatic detection and segmentation of teeth on panoramic radiographs
Hosntalab et al. (2010) [10]	Oral and maxillofacial radiology	Classification and numbering of teeth on multi-slice CT images
Miki et al. (2017) [11]	Oral and maxillofacial radiology	Classification of teeth on CBCT images
Lahoud et al. (2021) [12]	Oral and maxillofacial radiology	3D segmentation of tooth on CBCT images
Takahashi et al. (2020) [13]	Implantology	Identification of dental implants on panoramic radiographs
Hadj Saïd et al. (2020) [14]	Implantology	Identification of dental implants on a radiograph
Lee and Jeong (2020) [15]	Implantology	Identification of dental implants on periapical radiographs
Vranckx et al. (2020) [16]	Oral and maxillofacial radiology	Measurement of mandibular molar angulation
Park et al. (2019) [17]	Oral and maxillofacial radiology	Identification of cephalometric landmarks
Kunz et al. (2020) [18]	Oral and maxillofacial radiology	Identification of cephalometric landmarks
Hwang et al. (2020) [19]	Oral and maxillofacial radiology	Identification of cephalometric landmarks

(continued)

Table 1 (continued)

Study	Field	Application
Yu et al. (2020) [20]	Oral and maxillofacial radiology	Automated skeletal classification on cephalometric radiographs
Kök et al. (2019) [21]	Orthodontics	Automated determination of growth stage
Krois et al. (2019) [22]	Periodontology	Detection of periodontal bone loss on dental panoramic
Kim et al. (2019) [23]	Periodontology	Detection of periodontal bone loss on dental panoramic
Lee et al. (2018) [24]	Periodontology	Diagnose and predictions of bone loss on periapical radiographs
Chang et al. (2020) [25]	Periodontology	Diagnose and stage of periodontal bone loss on panoramic radiographs
Saghiri et al. (2012) [26]	Endodontics	Localization of the minor foramen
De Tobel et al. (2017) [27]	Forensic dentistry	Automated determination of lower third molar stage
Patil et al. (2020) [28]	Forensic dentistry	Gender determination based on mandibular morphometric parameters
Devito et al. (2008) [29]	Cariology	Diagnosis of proximal dental caries
Lee et al. (2018) [30]	Cariology	Detection and diagnosis of dental caries
Johari et al. (2017) [31]	Restorative dentistry	Detection of vertical root fractures
Fukuda et al. (2020) [32]	Restorative dentistry	Detection of vertical root fractures on panoramic radiographs
Aubreville et al. (2017) [33]	Estomatology	Automated classification of cancerous tissue in the oral cavity
Poedjiastoeti and Suebnukarn (2018) [34]	Estomatology	Diagnosis of jaw tumors
Thanathornwong (2018) [35]	Orthodontics	Assessment of the need for orthodontic treatment
Xie et al. (2010) [36]	Orthodontics	Assessment of the need for dental extractions prior to orthodontic treatment
Jung and Kim (2016) [37]	Orthodontics	Assessment of the need for dental extractions prior to orthodontic treatment
Choi et al. (2019) [38]	Orthodontics	Diagnose the need for orthognathic surgery
Zhang et al. (2018) [39]	Oral and maxillofacial surgery	Prediction of swelling following mandibular third molar extractions
Moghimi et al. (2012) [40]	Orthodontics	Prediction of the size of unerupted teeth
Chien-Hsun et al. (2009) [41]	Oral and maxillofacial surgery	Prediction of post-surgical changes in facial profile
Aliaga et al. (2015) [42]	Restorative dentistry	Prediction of longevity of dental restorations
Yamaguchi et al. (2019) [43]	Restorative dentistry	Prediction of debonding of CAD/CAM restorations
Alarifi et al. (2018) [44]	Implantology	Prediction of implant success rates
Cheng et al. (2015) [45]	Prosthodontics	Prediction of facial deformation with complete dentures
Wei et al. (2018) [46]	Prosthodontics	Prediction of ceramic recipes for dental color matching

Concluding Remarks

The development of automated tools for many specialties in dentistry is still in early phase; nevertheless, the results are promising. Some of the AI applications are summarized in Table 1. Most of the studies show the capacity of AI to perform human activities equally or better than specialists and in a shorter time without the concern of observer variability.

Currently, we are seeing the AI creeping into the field of dentistry as a modern dental assistant. Many companies have been working on the development of AI-based solutions in dentistry (Table 2). In the future, systems trained with large datasets for a wide range of clinical situations will be able to recognize a variety of features from inputs, such as intra- and extraoral radiographs, photos, tomographies, and digital models. This will allow automatic assessment of the patient's oral condition, diagnose, and

Table 2 Companies developing AI solutions in dentistry

Company	Country	Description
BubblesEndo	USA	AI dental startup transforming endodontic diagnosis. The product called “Endogenie” is an AI-based endodontic application that helps clinicians to organize their findings during examination, analyzes dental radiographs, and highlights its findings. The application also combines the imaging analysis with the results from clinical testing to provide the most likely clinical scenario Website: https://www.endogenie.ai
CellmatiQ	Germany	This company mission is to extend complex image analysis beyond human capabilities and automatically detect and classify structural patterns, optimize processing time and quality of image-based workflows, and improve visual assessments with objective and validated methods. Among the dental solutions available are modules for cephalometric analysis and image modality classification. Currently, the company is developing solutions for pathology indication in dental panoramic radiographs, caries detection, and age determination by cervical vertebral maturation Website: https://cellmatiq.com
Dental Intelligence	USA	The goals of this company are to help dentists to provide more and better care and to improve the team’s collaboration and performance, raising the overall health and profitability of the dental practice. Their dental solution is a module that connects to the practice management software and tracks the activities and analyzes, finds, and signalizes opportunities, making sure that time is saved, more patients are helped, and production is increased Website: https://www.dentalintel.com
Dental Monitoring	France	This French company wants to reinvent the patient experience and revolutionize the way oral healthcare is provided, expanding the dental practice into the virtual environment. Among their solutions are applications that monitor patients’ treatments, AI technology to generate ultra-realistic smile predictions, and automation of repetitive tasks Website: https://dental-monitoring.com
dentalXr.ai	Germany	With the combination of machine learning, software engineering, and expertise in dental research, this company is committed to support dentists in making the best possible diagnostic and treatment decisions for their patients. Their technology supports dentists in detecting pathologies and restorations in dental radiographs Website: https://dentalxr.ai
Denti.AI	Canada	Cloud-based artificial intelligence applied to interpret dental images and developed to work as a “second opinion,” helping in the identification of undiagnosed conditions and fast elaboration of dental chartings Website: https://www.denti.ai
Diagnocat	Russia	Platform that allows uploading, storing, and sharing of dental images with a built-in AI technology that automatically detects common dental conditions and pathologies Website: https://diagno.cat
glidewell.io	USA	Company that promises to improve patient care and increase profit by offering solutions that use machine learning technology and AI algorithms to create customized crown proposals and ensure accurate scans with color coding to show the missed areas Website: glidewell.io
Kapanu AG	Switzerland	This Swiss company aims to help the dental community achieve a high-quality and multidisciplinary services and products. Among them, there are software applications for visualizing the possibilities for aesthetic dental makeovers with augmented reality and preview of the potential orthodontic treatment outcome Website: https://kapanu.com
MMG Fusion	USA	The MMG Fusion solution connects to the dental practice management software and helps in finding available time for filling the chair, increasing office production and revenue, reducing staff overhead, and creating great patient’s experiences Website: https://mmgfusion.com

(continued)

Table 2 (continued)

Company	Country	Description
NovoDynamics	USA	NovoDynamics products transform large, disparate datasets into actionable information and insights that help organizations make better decisions with AI solutions addressing highly complex challenges. Among their products, there are solutions for radiographic analysis and enhancement and detection of anomalous claims submitted for dental insurances Website: https://www.novohealthdental.com
Orca Dental AI	Israel	The mission of this company is to provide dental practitioners the tools they need to make a breakthrough in patient care. It combines clinical experience and AI technologies to create diagnostic reports, treatment plan suggestions, and smart clinical predictions. Among their products, there are automated cephalometric analysis, airway analysis, mandibular nerve canal identification in CBCTs, and caries detection Website: https://www.orca-ai.com
Overjet	USA	This company incubated at the Harvard Innovation Labs is dedicated to solving the most challenging problems in dentistry. They have modules for insurance claims review, automation of administrative tasks, and identification of dental anatomy, oral diseases, and quality of restorations Website: https://www.overjet.ai
Pearl	USA	Founded on the notion that AI can be the dental assistant, Pearl aims to usher in a new wave of AI-powered tools, such as second opinion on every radiograph, management of dental practices, delimitation of indirect restoration margins, and automated insurance claims reviews Website: https://hellopearl.com
Relu BV	Belgium	Relu is a young and driven company that allies AI engineering and research expertise to develop clinical applications. Among their solutions, there are modules for automatic segmentation of teeth on CBCT scans, calculation of molar angulation to predict third molar eruption on dental panoramics, detection of the inferior alveolar nerve, and segmentation of the pharynx, mandible, and skull Website: https://relu.eu
Simplifeye	USA	Company focused on providing cutting-edge software technology and world-class service. They offer online solutions to reach patients anytime and anywhere and to access their information available in the cloud Website: https://simplifeye.co
VideaHealth	USA	The Boston-based company provides AI-driven tools for image analysis, diagnosis, treatment planning, and insurer claims processing Website: https://www.vedea.ai

suggest the best treatment options with the most optimal prognosis. Further, it will pave the way for predicting the outcomes based on the particular characteristics of each case.

References

- Khanagar SB, Al-Ehaideb A, Maganur PC, Vishwanathaiah S, Patil S, Baeshen HA, et al. Developments, application, and performance of artificial intelligence in dentistry - a systematic review. *J Dent Sci.* 2021;16(1):508–22.
- Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. *J Dent Res.* 2020;99(7):769–74.
- Joda T, Bornstein MM, Jung RE, Ferrari M, Waltimo T, Zitzmann NU. Recent trends and future direction of dental research in the digital era. *Int J Environ Res Public Health.* 172020.
- Leite AF, Vasconcelos KF, Willems H, Jacobs R. Radiomics and machine learning in Oral healthcare. *Proteomics Clin Appl.* 2020;14(3):e1900040.
- Jaskari J, Sahlsten J, Järnstedt J, Mehtonen H, Karhu K, Sundqvist O, et al. Deep learning method for Mandibular Canal segmentation in dental cone beam computed tomography volumes. *Sci Rep.* 2020;10(1):5842.
- Zhang K, Wu J, Chen H, Lyu P. An effective teeth recognition method using label tree with cascade network structure. *Comput Med Imaging Graph.* 2018;68: 61–70.
- Chen H, Zhang K, Lyu P, Li H, Zhang L, Wu J, et al. A deep learning approach to automatic teeth detection

- and numbering based on object detection in dental periapical films. *Sci Rep.* 2019;9(1):3840.
8. Tuzoff DV, Tuzova LN, Bornstein MM, Krasnov AS, Kharchenko MA, Nikolenko SI, et al. Tooth detection and numbering in panoramic radiographs using convolutional neural networks. *Dentomaxillofac Radiol.* 2019;48(4):20180051.
 9. Leite AF, Gerven AV, Willems H, Beznik T, Lahoud P, Gaêta-Araujo H, et al. Artificial intelligence-driven novel tool for tooth detection and segmentation on panoramic radiographs. *Clin Oral Investig.* 2020.
 10. Hosntalab M, Aghaeizadeh Zoroofi R, Abbaspour Tehrani-Fard A, Shirani G. Classification and numbering of teeth in multi-slice CT images using wavelet-Fourier descriptor. *Int J Comput Assist Radiol Surg.* 2010;5(3):237–49.
 11. Miki Y, Muramatsu C, Hayashi T, Zhou X, Hara T, Katsumata A, et al. Classification of teeth in cone-beam CT using deep convolutional neural network. *Comput Biol Med.* 2017;80:24–9.
 12. Lahoud P, EzEldeen M, Beznik T, Willems H, Leite A, Van Gerven A, et al. Artificial intelligence for fast and accurate 3D tooth segmentation on CBCT. *J Endod.* 2021.
 13. Takahashi T, Nozaki K, Gonda T, Mameno T, Wada M, Ikebe K. Identification of dental implants using deep learning-pilot study. *Int J Implant Dent.* 2020;6(1):53.
 14. Hadj Saïd M, Le Roux MK, Catherine JH, Lan R. Development of an artificial intelligence model to identify a dental implant from a radiograph. *Int J Oral Maxillofac Implants.* 2020;36(6):1077–82.
 15. Lee JH, Jeong SN. Efficacy of deep convolutional neural network algorithm for the identification and classification of dental implant systems, using panoramic and periapical radiographs: a pilot study. *Medicine (Baltimore).* 2020;99(26):e20787.
 16. Vranckx M, Van Gerven A, Willems H, Vandemeulebroucke A, Ferreira Leite A, Politis C, et al. Artificial intelligence (AI)-driven molar angulation measurements to predict third molar eruption on panoramic radiographs. *Int J Environ Res Public Health.* 2020;17(10).
 17. Park JH, Hwang HW, Moon JH, Yu Y, Kim H, Her SB, et al. Automated identification of cephalometric landmarks: part 1-comparisons between the latest deep-learning methods YOLOV3 and SSD. *Angle Orthod.* 2019;89(6):903–9.
 18. Kunz F, Stellzig-Eisenhauer A, Zeman F, Boldt J. Artificial intelligence in orthodontics : evaluation of a fully automated cephalometric analysis using a customized convolutional neural network. *J Orofac Orthop.* 2020;81(1):52–68.
 19. Hwang HW, Park JH, Moon JH, Yu Y, Kim H, Her SB, et al. Automated identification of cephalometric landmarks: part 2- might it be better than human? *Angle Orthod.* 2020;90(1):69–76.
 20. Yu HJ, Cho SR, Kim MJ, Kim WH, Kim JW, Choi J. Automated skeletal classification with lateral Cephalometry based on artificial intelligence. *J Dent Res.* 2020;99(3):249–56.
 21. Kök H, Acilar AM, İzgi MS. Usage and comparison of artificial intelligence algorithms for determination of growth and development by cervical vertebrae stages in orthodontics. *Prog Orthod.* 2019;20(1):41.
 22. Krois J, Ekert T, Meinholt L, Golla T, Kharbot B, Wittmeier A, Dörfer C, Schwendicke F. Deep learning for the radiographic detection of periodontal bone loss. *Sci Rep.* 2019;9(1):8495.
 23. Kim J, Lee HS, Song IS, Jung KH. DeNTNet: deep neural transfer network for the detection of periodontal bone loss using panoramic dental radiographs. *Sci Rep.* 2019;9(1):17615.
 24. Lee JH, Kim DH, Jeong SN, Choi SH. Diagnosis and prediction of periodontally compromised teeth using a deep learning-based convolutional neural network algorithm. *J Periodontal Implant Sci.* 2018;48(2):114–23.
 25. Chang HJ, Lee SJ, Yong TH, Shin NY, Jang BG, Kim JE, Huh KH, Lee SS, Heo MS, Choi SC, Kim TI, Yi WJ. Deep learning hybrid method to automatically diagnose periodontal bone loss and stage periodontitis. *Sci Rep.* 2020;10(1):7531.
 26. Saghiri MA, Garcia-Godoy F, Gutmann JL, Lotfi M, Asgar K. The reliability of artificial neural network in locating minor apical foramen: a cadaver study. *J Endod.* 2012;38(8):1130–4.
 27. De Tobel J, Radesh P, Vandermeulen D, Thevissen PW. An automated technique to stage lower third molar development on panoramic radiographs for age estimation: a pilot study. *J Forensic Odontostomatol.* 2017;35(2):42–54.
 28. Patil V, Vineetha R, Vatsa S, Shetty DK, Raju A, Naik N, et al. Artificial neural network for gender determination using mandibular morphometric parameters: a comparative retrospective study. *Cogent Eng.* 2020;7(1):1723783.
 29. Devito KL, de Souza BF, Filho WNF. An artificial multilayer perceptron neural network for diagnosis of proximal dental caries. *Oral Surg Oral Med Oral Pathol Oral Radiol Endodontol.* 2008;106(6):879–84.
 30. Lee J-H, Kim D-H, Jeong S-N, Choi S-H. Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. *J Dent.* 2018;77:106–11.
 31. Johari M, Esmaeili F, Andalib A, Garjani S, Saberkari H. Detection of vertical root fractures in intact and endodontically treated premolar teeth by designing a probabilistic neural network: an ex vivo study. *Dentomaxillofac Radiol.* 2017;46(2):20160107.
 32. Fukuda M, Inamoto K, Shibata N, Ariji Y, Yanashita Y, Kutsuna S, et al. Evaluation of an artificial intelligence system for detecting vertical root fracture on panoramic radiography. *Oral Radiol.* 2020;36(4):337–43.
 33. Aubreville M, Knipfer C, Oetter N, Jaremenko C, Rodner E, Denzler J, et al. Automatic classification of cancerous tissue in Laserendomicroscopy images of the Oral cavity using deep learning. *Sci Rep.* 2017;7(1):11979.
 34. Poedjiastoeti W, Suebnukarn S. Application of convolutional neural network in the diagnosis of jaw Tumors. *Healthcare Inform Res.* 2018;24(3):236–41.

35. Thanathornwong B. Bayesian-based decision support system for assessing the needs for orthodontic treatment. *Healthcare Inform Res.* 2018;24(1):22–8.
36. Xie X, Wang L, Wang A. Artificial neural network modeling for deciding if extractions are necessary prior to orthodontic treatment. *Angle Orthod.* 2010;80(2):262–6.
37. Jung SK, Kim TW. New approach for the diagnosis of extractions with neural network machine learning. *Am J Orthod Dentofac Orthop.* 2016;149(1):127–33.
38. Choi HI, Jung SK, Baek SH, Lim WH, Ahn SJ, Yang IH, et al. Artificial intelligent model with neural network machine learning for the diagnosis of orthognathic surgery. *J Craniofac Surg.* 2019;30(7):1986–9.
39. Zhang W, Li J, Li ZB, Li Z. Predicting postoperative facial swelling following impacted mandibular third molars extraction by using artificial neural networks evaluation. *Sci Rep.* 2018;8(1):12281.
40. Moghim S, Talebi M, Parisay I. Design and implementation of a hybrid genetic algorithm and artificial neural network system for predicting the sizes of unerupted canines and premolars. *Eur J Orthod.* 2012;34(4):480–6.
41. Lu C-H, Ko EW-C, Liu L. Improving the video imaging prediction of postsurgical facial profiles with an artificial neural network. *J Dental Sci.* 2009;4(3):118–29.
42. Aliaga IJ, Vera V, De Paz JF, García AE, Mohamad MS. Modelling the longevity of dental restorations by means of a CBR system. *Biomed Res Int.* 2015;2015:540306.
43. Yamaguchi S, Lee C, Karaer O, Ban S, Mine A, Imazato S. Predicting the Debonding of CAD/CAM composite resin crowns with AI. *J Dent Res.* 2019;98(11):1234–8.
44. Alarifi A, AlZubi AA. Memetic search optimization along with genetic scale recurrent neural network for predictive rate of implant treatment. *J Med Syst.* 2018;42(11):202.
45. Cheng C, Cheng X, Dai N, Jiang X, Sun Y, Li W. Prediction of facial deformation after complete denture prosthesis using BP neural network. *Comput Biol Med.* 2015;66:103–12.
46. Wei J, Peng M, Li Q, Wang Y. Evaluation of a novel computer color matching system based on the improved Back-propagation neural network model. *J Prosthodont.* 2018;27(8):775–83.



Artificial Intelligence in Gastroenterology

66

Inga Strümke, Steven A. Hicks, Vajira Thambawita, Debesh Jha,
Sravanthi Parasa, Michael A. Riegler, and Pål Halvorsen

Contents

Introduction	920
GI Endoscopy	921
Existing Methods	922
Hand-Crafted-Feature-Based Approaches	923
Deep Learning-Based Approaches	923
Unsupervised and Semi-supervised Approaches	924
Example Results	925
Open Issues and Ongoing Research	925
Limited Data Availability	926
Generalizability	928
Metrics and Evaluation	928
Automatic Report Generation	929
Explainability	930

I. Strümke · M. A. Riegler

SimulaMet, Oslo, Norway

e-mail: inga@simula.no; michael@simula.no

S. A. Hicks · V. Thambawita · P. Halvorsen (✉)

SimulaMet, Oslo, Norway

Department of Computer Science, Oslo Metropolitan

University, Oslo, Norway

e-mail: steven@simula.no; vajira@simula.no;
paalh@simula.no; pallh@oslomet.no

D. Jha

SimulaMet, Oslo, Norway

Department of Computer Science, UIT The Arctic

University of Norway, Oslo, Norway

e-mail: debesh@simula.no

S. Parasa

Department of Gastroenterology, Swedish Medical Group,
Seattle, WA, USA

Competitions and Challenges	931
Clinical Verification and Emerging Commercial Systems	931
Summary and Conclusions	932
References	933

Abstract

The holy grail in endoscopy examinations has for a long time been assisted diagnosis using Artificial Intelligence (AI). Recent developments in computer hardware are now enabling technology to equip clinicians with promising tools for computer-assisted diagnosis (CAD) systems. However, creating viable models or architectures, training them, and assessing their ability to diagnose at a human level, are complicated tasks. This is currently an active area of research, and many promising methods have been proposed. In this chapter, we give an overview of the topic. This includes a description of current medical challenges followed by a description of the most commonly used methods in the field. We also present example results from research targeting some of these challenges, and a discussion on open issues and ongoing work is provided. Hopefully, this will inspire and enable readers to future develop CAD systems for gastroenterology.

Keywords

Gastrointestinal endoscopy · Artificial Intelligence · Neural Networks · Hand-crafted features · Anomaly detection · Semantic segmentation · Performance

Introduction

Numerous abnormal mucosal findings, ranging from minor annoyances to highly lethal diseases, can be found in the human Gastrointestinal (GI) tract. For example, according to the International Agency for Research on Cancer, about 3.5 million luminal GI (esophageal, stomach, colorectal) cancers are detected yearly in the world [41]. These cancers represent a substantial health

challenge for society, with a mortality rate of about 63–65%, resulting in around 2.2 million deaths per year [19, 41]. Overall, Colorectal cancer (CRC) is the third most common cause of cancer mortality for women and men combined [104], and the other most frequently occurring GI cancers are stomach, liver, pancreatic, and esophageal cancers [18].

For diagnosis and treatment of GI diseases, GI endoscopy is the gold-standard procedure used to examine the tract for anomalies, and to a certain extent, the GI diseases may be prevented by improved endoscopic performance and high quality systematic screening in high incidence areas [19]. However, despite the substantial technical improvement of endoscopes over the last two decades, a major limitation of the endoscopic examinations is the endoscope operator variation, depending on the procedural skill, perceptual factors, personality characteristics, experience, knowledge, and attitude deficits [34]. This translates to a substantial inter-observer variation in the detection and assessment of mucosal lesions [64, 108]. This causes, for example, an average 20% polyp miss-rate during colonoscopies [52]. All these factors could potentially, to some extent, be alleviated by substantial educational efforts, but not eliminated [88].

In this context, assisted diagnosis using computers has for a long time been a holy grail. Developments in computer hardware have enabled computationally demanding yet promising technologies like AI, more specifically its sub-field Machine Learning (ML), to provide the clinicians with potentially highly accurate and efficient Computer aided diagnosis (CAD) systems, giving healthcare professionals the tools needed to provide quality care at a large scale [86, 102]. At its core, machine learning involves using algorithms to parse data, learn from it, and then make predictions, in the medical domain this means detect,

segment, assess or classify a disease. However, there exist several issues which need to be addressed, both for creating and improving automated diagnosis algorithms. Developing and assessing a computer's ability to diagnose at a human level are complicated tasks, and a potential success depends on various factors which goes beyond simply determining the accuracy of an algorithm. These challenges have been an active area of research for about a decade, and a large number of promising results have been published.

In this chapter, we describe current challenges on the way towards effective computer-based digital assistant systems. In particular, we focus on GI endoscopy. We provide examples of proposed methods and tools employing various techniques, identify current challenges, and give hints for future development and assessment of CAD systems.

GI Endoscopy

To examine the esophagus, stomach, duodenum (upper GI), and the large bowel and rectum (lower GI), a long, flexible tube is inserted into the mouth and rectum, respectively. A tiny video camera at the tip of the tube allows the doctor to view inside of the GI tract in real-time, where findings, as depicted in Fig. 1a and b can be found.

The small bowel is, due to its anatomical location, less accessible for inspection by such flexible endoscopes. To easier access these areas of the GI tract, Video Capsule Endoscopy (VCE) [22] has been introduced as an alternative examination method [25]. A VCE consists of a small capsule containing one or more wide-angle cameras. The capsule is swallowed by the patient, and it captures a video as it moves through the GI tract. The video is extracted, and a medical expert assesses it in a potentially tedious and time-consuming process after the procedure, searching for findings like the ones shown in Fig. 1c.

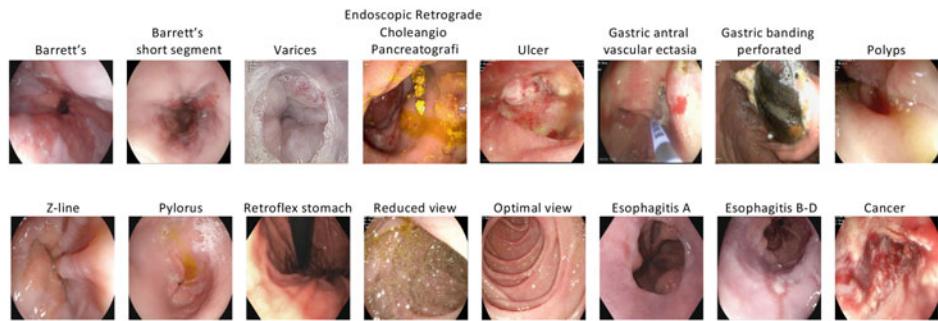
Even though these examination procedures allow clinicians to detect GI anomalies, there is still ample scope for improvements. Looking at the possible findings depicted in Fig. 1, it is obvious that it can be hard to detect and classify the various anomalies potentially found in the various

parts of the GI tract, either live during a gastroscopy or colonoscopy, or in a post-analysis of the VCE video. Moreover, there are large operator variations and anomaly miss-rates reported for both regular endoscopies [34, 52, 64, 108] and capsule endoscopies [20, 88].

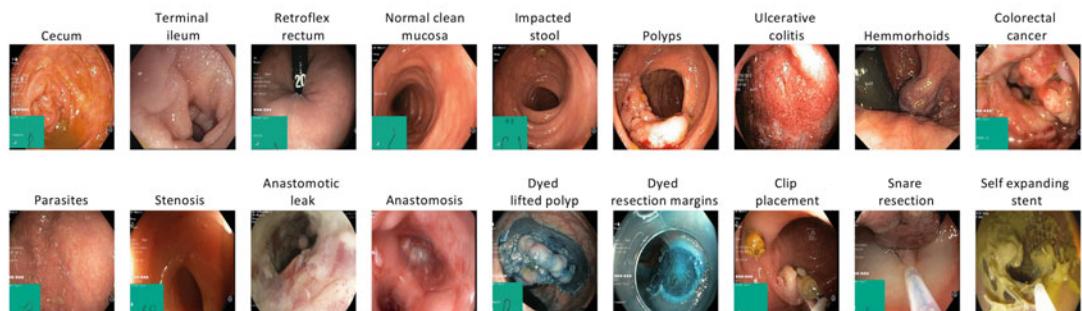
Hence, the hope is that automated analysis can assist medical experts in real-time anomaly detection, removing variations and increasing detection rates. Moreover, analyzing hours of VCE video, there is also a large potential in saving medical expert time, by analyzing the 4–12 h long videos in a few minutes by a fast computer, compared to the usual 45–60 min error-prone, fast-forward analysis performed by medical personnel today. From an analysis point of view, there are two important requirements for such CAD systems:

1. *High detection or segmentation performance* in the analysis is important in order to address the large human miss-rates and variabilities. It is often measured in terms of metrics like precision, sensitivity (recall), specificity, accuracy, F1 score, Matthews correlation coefficient (MCC) or similar [98]. This requirement aims at finding all anomalies correctly, i.e., detecting all findings without false positives or negatives. A more detailed discussion on metrics is given in section “Metrics and Evaluation.”
2. An often neglected requirement is *fast processing* in order to give real-time feedback during the endoscopy examination, or in the case of VCE, higher scale of the analysis and a faster feedback on the same amount of processing resources.

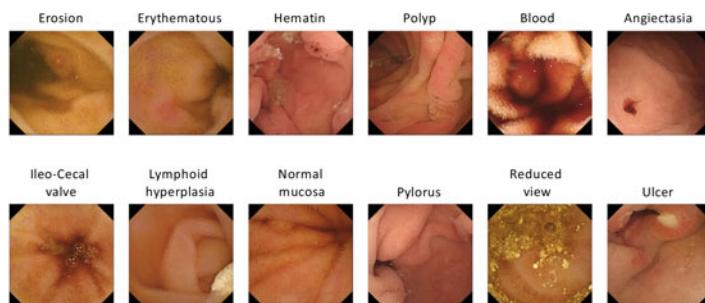
Furthermore, in order to be deployable in a clinical environment, all components need to be integrated in a pipeline capturing videos or frames from the endoscopy equipment, via an automatic analysis, to give the clinicians a visual feedback (and potentially also assisting in generating an examination report according to medical standards). The system must also be easily integrated into and usable with the current examination procedures, and of course, the various components must meet the medical privacy and security regulations.



(a) Upper GI tract, esophagus, stomach (Gastroscopy).



(b) Lower GI tract, large bowel (Colonoscopy).



(c) Lower GI tract, small bowel (Capsule endoscopy).

Fig. 1 Examples of various findings in the GI tract including anatomical landmarks, pathological findings, normal mucosa, therapeutic interventions and medical instruments [14, 96]

Existing Methods

As mentioned above, a large number of algorithms and models for automated analysis of GI video and images have already been proposed. In this respect, when we discuss CAD systems for the GI tract today, people often interchangeably

talk about detection, localization and segmentation. Here, we therefore first try to distinguish between the terms as follows:

- *Detection* is the operation of detecting whether an image belongs to a certain classification or not. This can be a binary “yes or no” for



Fig. 2 Various ways of indicating a finding (*left*: detection “just” showing the image; *center*: full segmentation mask showing in white all pixels part of the finding; and *right*: bounding box making a rectangle around the finding)

questions whether the image or video frame contains a polyp or not. It also includes systems that *classifies* the input into multiple classes.

- *Localization* is to point into the image where the object is located, e.g., using some type of point markers or making a bounding box around the object of interest.
- *Segmentation* is yet another step further where one determines pixel-wise whether the pixel belongs to a finding or not, e.g., generating an exact segmentation mask of the finding.

Figure 2 shows an example of detection, localization and detection. As localization is often mixed into both detection and segmentation, we here focus on detection and segmentation.

Hand-Crafted-Feature-Based Approaches

Automatic detection of GI anomalies has been a topic of research long before the success of AI and deep neural networks, using what is nowadays often called traditional computer vision and ML methods, as found in libraries such as *OpenCV* [16] and *LIRE* [69]. Already in 1998, Krishnan et al. [59] proposed detecting polyps using shape-features in a curvature analysis. In the subsequent decade, various approaches using a mix of shape, edge, texture, and color features appeared. For example, Alexandre et al. [2] detected polyps using a support vector machine (SVM) on color patterns. Further, using SVMs, Ameling et al. [5]

combined texture and colors, and Park et al. [75] used shape and texture features in a conditional random field classifier.

Two more recent approaches using hand-crafted techniques are Polyp-Alert [111] and EIR [85], where the authors also measured analysis time, with the goal of being able to give real-time feedback during the examination. The Polyp-Alert [111] system combines edge and texture features. The polyp edge detection algorithm mainly relies on edge features obtained from the part-based multi-derivative edge cross-section profile [110]. The EIR [85] system combines a content-based similarity search with statistical classifiers from the training data. A large number of image features are tested [87], ending up with a combination of the joint composite descriptor feature and the Tamura features, due to a good trade-off between the precision and sensitivity (recall), and the speed of the algorithm. A search-based classifier is then used to determine if an image contains a finding of a certain class.

A detailed overview containing earlier example approaches can be found in [85, 111]. However, lately, deep learning approaches have outperformed these hand-crafted approaches and replaced them entirely.

Deep Learning-Based Approaches

Already in 2001, Karkanis et al. [53] aimed for the detection of lesions in endoscopic video using textural descriptors on the wavelet domain supported by artificial neural network

architectures, albeit not using deep architectures. Such early approaches were tested on tiny data sets, in this case 8 images [53]. More recent approaches are usually based on deep learning architectures where Convolutional Neural Networks (CNNs) are clearly the most popular ones.

Where hand-crafted features rely on extracting predefined properties of an image, such as color, texture, or shape, CNNs are neural network architectures using convolutions and pooling operations to automatically learn which features are most relevant. CNNs perform well on many different tasks like image classification, object detection in images, and image generation [56]. Although they are mostly used for image analysis, they have also proven useful in timeseries research and video analysis. In medicine, architectures like U-Net [89] have shown promising results in areas like cardiology, colonoscopy, and radiology [74, 119, 122]. This also includes gastroenterology, where CNNs are currently state-of-the-art for analyzing colonoscopy videos. The most common application is the detection and segmentation of polyps, where many CNN-based approaches have shown excellent results [17, 50, 114]. These approaches have expanded to other findings as well, like detecting and segmenting ulcers [31]. Furthermore, due to limited access to medical image and video data, most approaches use transfer learning. In transfer learning, pre-trained models are used as a starting point, and refined for the given data set by retraining with some layers trainable and some frozen [82].

An automated CAD system for the GI endoscopic image segmentation is a step further than providing “just” detection of anomalies. A predicted segmentation mask (see Fig. 2) can help point out the area of interest in the images (frames) that need to be further examined. However, making such perpixel predictions is also a more complex task. In this respect, there has been a considerable amount of work done so far, especially targeting polyps [32, 45, 47, 48, 50, 71, 81, 100, 109], artifacts [3], and endoscopic instruments [90]. In general, CNN-based approaches perform well with the larger polyps. However, still the major challenges issues in the field are

related to adenomatous polyps or small and flat polyps. Recent studies are targeting smaller polyps [50, 63]; however, it is yet an open-challenge to solve.

Unsupervised and Semi-supervised Approaches

The above presented approaches fall into the category of supervised learning, meaning that we train the models on a data set with an existing ground truth. In this section, we give a glance at newly emerging unsupervised and semi-supervised methods.

Generative Adversarial Networks (GANs), which were introduced by Goodfellow et al. [30] in 2014, are becoming increasingly popular in the medical domain for generating synthetic data. Different advancements to the original GAN architecture, such as conditional GAN [72], pix2pix [43], CycleGAN [123], Style-GANs [54, 55], to mention a few, present different methods, ranging from domain transformation to high definition image generation. ML researchers in the medical domain can use GAN models to generate synthetic data to tackle challenges related to privacy, data deficiency, and data annotation. For example, Younghak et al. [93] use a conditional GAN architecture to generate synthetic polyp images to improve the performance of a deep learning system detecting polyps in the colon. This methodology is still in its early stages, and it has yet to be shown to which extent generated data can replace real data and help to improve performance and shareability.

Another emerging method in the field of medical image analysis, is semi-supervised learning. Here, the goal is to learn from a small set of labelled data combined with a larger amount of unlabeled data. Examples include [7, 67, 70, 116]. These models produce promising results, and could also help overcome the challenge of insufficient labelled data faced by many data-hungry methods. However, these approaches still struggle with challenges such as low accuracy and high entropy during early stages of the training

process. The models are also regularized towards high entropy predictions, making it hard to achieve a high accuracy [117, 120]. It will be interesting to see whether these challenges can be overcome, and how useful the results will prove to be in the medical domain.

Example Results

High detection or segmentation rates are important in order to be clinically relevant, and the typical way the performance is compared. However, due to factors like different data sets and different equipment, the pure numbers cannot be directly compared. Still, to give some indications of the state-of-the-art performance, we give a set of, by far from complete, examples using standard metrics like precision, sensitivity (recall), specificity, accuracy, F1 score and MCC for detection; and Dice similarity coefficient (DSC), Intersection over Union (IoU), precision and sensitivity for segmentation. A substantial overview of existing approaches can be found in [61], containing 138 different studies. An explanation of the different metrics is given in Table 1 and further discussed in section “Metrics and Evaluation.” Another source for exploring and comparing different approaches are the popular GI detection,

classification and segmentation challenges discussed in section “Competitions and Challenges.”

A selection of performance examples are given in Table 2. Looking at the numbers, we see that in the specific tested cases, the computer should be at the level of the best experts with scores above 90%, i.e., potentially being a helpful digital assistant during a GI endoscopy examination. Likewise, example results for lesion segmentation are provided in Table 3, and the numbers are again encouraging in terms of proving that the used models could be of use in a medical setting. However, while the results achieved are promising, there are still several open challenges, including generalizability, overfitting, cross data set testing and explainability of the results. Moreover, as indicated in the Tables, hardly any existing research report the speed of the system, meaning that it is hard to assess the system’s capability to provide a live analysis in the clinic.

Open Issues and Ongoing Research

Despite impressive results presented in many of the published papers, even exceeding what are reported as average detection rates from clinicians, there are still challenges and open issues.

Table 1 List of commonly used metrics. To define each metric, TP, FP, TN, and FN represent true positives, false positives, true negatives, and false negatives, respectively

Formula	Description
$\text{accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}}$	Rate of correct classification. Ratio between correctly classified samples and all samples.
$\text{precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$	Proportion of retrieved samples which are relevant. Ratio between correctly classified positive samples and all samples classified as positive.
sensitivity (also known as recall) $= \frac{\text{TP}}{\text{TN}+\text{FP}}$	Proportion of relevant samples which are retrieved. Ratio between correctly classified positive samples and all positive samples.
$\text{specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}}$	Negative class sensitivity. Ratio between correctly classified negative samples and all negative samples.
$\text{F1} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}$	Harmonic mean of the precision and sensitivity (recall).
$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})d(\text{TP}+\text{FP}+\text{FN})}}$	Pearson’s correlation coefficient [23] for binary classification.
IoU (also known as Jaccard) $= \frac{\text{TP}}{\text{TP}+\text{FP}+\text{FN}}$	Similarity between sets from the size of the intersection divided by the size of the union.
$\text{DSC} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}$	Quotient of similarity of two sets. Semi-metric as it doesn’t satisfy the triangle inequality. Related to the IoU via $\frac{S}{2-S}$.

Table 2 Examples of **detection** performance from different approaches. The results show promising performance with numbers above 90%. Unfortunately, speed is not commonly reported

Paper/system	Data set used	Sensitivity (recall)	Specificity	Accuracy	Precision	F1	MCC	Speed (fps)
Boughorbel [17]	MICCAI-challenge data sets	86.3	–	–	73.6	–	–	–
Kundu [60]	30 Own data set	95.2	98.3	97.9	88.4	–	–	–
Cho [21]	Seoul National University Hospital	>87	–	>93	–	–	–	–
Ghosh [29]	VCE videos data set	99.4	99.2	97.9	95.8	–	–	–
Bell [8]	CTC generating 4000 images per patients	89.8	75.5	–	–	–	–	–
Pogorelov [76]	Kvasir	83.9	98.5	97.2	84.1	85.6	82.8	46
Billah [13]	Colonoscopy & Endoscopy vision data set	98.7	98.2	98.3	–	–	–	–
Thambawita [99]	Kvasir	95.8	99.7	95.8	95.9	95.8	95.3	29

Table 3 Some examples of different **segmentation** approaches applied to different data sets. We can clearly see that the performance overall is quite promising (with all

metrics in the range of 70 to 95). Speed is unfortunately not commonly reported

Paper/system	Data set used	DSC	IoU (Jaccard)	Sensitivity (recall)	Precision	Speed (fps)
U-Net [89]	MICCAI-PhC-U373	–	92.0	–	–	–
PraNet [26]	CVC-ClinicDB	89.9	84.0	–	–	–
PolypSegNet [71]	CVC-ClinicDB	91.5	86.2	91.1	96.2	–
ResUNet++ [50]	CVC-ClinicDB	79.6	79.6	70.2	87.9	–
PraNet [26]	Kvasir-SEG	89.8	84.0	–	–	–
PolypSegNet [71]	Kvasir-SEG	88.7	82.5	84.5	91.7	–
ResUNet++ [50]	Kvasir-SEG	81.3	79.3	70.6	87.7	–
Double-UNet [45]	CVC-ClinicDB	92.4	86.1	84.6	96.0	–

First, for example, Thambawita et al. [98] presented the issue of overfitting to specific data sets and a lack of generalizability. This means that a model that performs well on one data set may not perform at all on another. Furthermore, like other deep neural networks, CNNs are black boxes, and it is not easy to understand why one input gives a particular result. There is also a lack of large open data sets that contain annotations for uncommon abnormalities and rarely documented findings to support data-hungry algorithms like CNNs. Here, we elaborate on a few of these open issues.

Limited Data Availability

Available medical data is scarce. However, modern deep learning approaches usually require a lot of data to perform well, and often, the more variations in the data, the better the model gets, especially for supervised learning models. Table 4 shows the data sets available in the field of GI endoscopy. Evidently, the number of images used for training and testing is small when compared to the data set from the natural images. This is because it is difficult to obtain data from the medical domain. The data is often protected and

Table 4 Summary of available endoscopic data sets. A further discussion about data sets are found in [14, 96]

Data set	Findings	Location	#Images	#Videos	Bounding box?	Segmentation mask?	Input size	VCE data?	Endoscopic device
Kvasir [77]	various	↑↓	8,000	—	—	—	variable	—	†
Nerthus [78]	stool (cleanness)	↓	5,525	—	—	—	720 × 576	—	†
HyperKvasir [14]	various	↑↓	110,079	373	—	—	variable	—	†
KvasirInstrument [46]	instruments	↓	590	—	✓	✓	variable	—	†
Kvasir-SEG [49]	polyps	↓	1,000	—	✓	✓	variable	—	†
ASU-Mayo [97]	polyps	↓	18,781	—	—	—	variable	—	—
CVC-ClinicDB [10]	polyps	↓	612	—	—	—	384 × 288	—	‡*
CVC-ColonDB [11]	polyps	↓	380	—	—	—	574 × 500	—	—
ETIS Larib Polyp DB [95]	polyps	↓	196	—	—	—	1225 × 966	—	‡*
SUN Colonoscopy Video DB [73]	polyps	↓	158,690	—	✓	—	1080 × 1240	—	?
CVCVideoClinicDB [6, 12]	polyps	↓	11,954	—	—	—	384 × 288	—	—
CAD-CAP [65]	various		25,000	—	—	—	—	✓	—
KID [58]	various		2,371	47	—	—	—	✓	—
KvasirCapsule [96]	various	↓	4,820,739	118	✓	—	variable	✓	•

Location: ↑ = upper GI ↓ = lower GI

Device: † = ScopeGuide, Olympus ‡ = Olympus Q160ALandQ165L

◦ = Exera IIvideoprocessor • = Olympus EC-S10 endocapsule

unavailable due to legal restrictions and lack of medical personnel for the tedious process of manually extracting and labeling training data. This calls either for better data sharing processes and culture, or methods more capable of handling small amounts of data.

This gives rise to several basic challenges: The amount of data is too small to train a robust model, and the presented results might appear deceptively good due to overfitting. Moreover, it is hard to compare results if all experiments are performed on different data, and practically impossible to reproduce them. Thus, it is almost impossible to conclude whether one model is better than another. We must therefore aim for more and open data sets. Table 4 contains an overview of known available data sets at the time of writing, making a good starting point for future experiments. Still, more data is needed, especially data containing pathological outcomes.

Generalizability

One of the open issues in the field is the GI endoscopy is the generalizability of ML models, i.e., their ability to perform well on previously unseen data regardless of source, equipment, etc. Such data can be from either the same distribution as the model was trained on, or from a different distribution. Which of the two a new data sample represents, is not always clear [101, 103]. Although some recent studies address generalizability of ML models for polyp classification [45, 112], this must be addressed for any model or system to be deployed into clinical practice.

Evaluating whether a model is reliable for real world use also requires cross data set testing, to avoid accepting a model which coincidentally works well on one specific set of data. The model developers should in general not have access to the final test data, to avoid bias during testing and development. This process, known as data blinding, is an important tool in many fields of research, including medicine [80]. Ideally, the

model should be tested for robustness on data collected separately from the data used during model development and testing.

Furthermore, distinction should be made between data annotated by medical experts, referred to as *soft ground truth*, and data labelled based on a medical test, referred to as *hard ground truth*, e.g., pathological examination of a polyp. The quality of soft ground truth data is limited by how well the medical annotator is trained, and such data is most useful for training models intended to automate processes. On the other hand, data with hard ground truth labels can also be used for automating processes, with the added benefit of avoiding annotator error or bias into the model, but it can furthermore be used for obtaining new knowledge. Note that while, as mentioned above, annotating each image is time consuming, collecting hard ground truth data is even more demanding, resulting in a scarcity of such data sets.

In current endoscopy practices, different hospitals use different endoscope system for diagnosis and therapy. The most common globally available endoscope systems are Olympus (Japan), Pentax 90i series (Japan), Fujinon (Japan), and Karl Storz (Germany) [57]. Moreover, different medical institutes have different protocols. Therefore, designing generalizable CAD systems is essential for performing well on a variety of institutes. Such systems should always be tested on several data sets. Discussions regarding challenges and advantages associated with cross-dataset testing can be found in [98].

Metrics and Evaluation

Evaluating performance is an important step when creating models for clinical use, and depends strongly on the choice of metric. As shown in Table 1, commonly used metrics are precision, sensitivity (recall), specificity, accuracy and F1 score. Some papers also report AUROC (area under the receiver operating characteristics). There are several reasons for going beyond the aforementioned metrics [98]. One challenge

frequently encountered in association with medical data sets, is their tendency to be imbalanced between classes, often having far more normal images than images with lesions. Because of this, certain metrics can provide an overly optimistic impression of the actual performance. For instance, a binary classifier can achieve a high accuracy on a data set containing few negative instances, by assigning all instances to the positive class. The AUROC is also known to be deceptive for imbalanced classification [91]. In such cases, the correlation coefficient between the true and predicted classes can be more informative [15], although no single metric is universally informative or suited for any imbalanced data problem. Moreover, for detection purposes, it is also a question whether one report per-frame performance, i.e., giving a decision for every frame in the video, or per-lesion, i.e., giving a correct prediction for at least one of the frames in the video sequence. Looking at the results from a technical point of view, a per-frame analysis of often desired, but from the medical point of view, a per-lesion analysis is often sufficient to notify the clinician of the finding once.

For segmentation performance, commonly used metrics are DSC and the IoU, also known as the Jaccard index. In clinical use, medical experts are usually interested in pixel-wise detail information about the potential lesion. DSC and IoU can be used to compare the pixel-wise similarity between the predicted segmentation maps and the ground truth. In addition, precision and sensitivity are used to evaluate under-segmentation or over-segmentation, where under-segmentation implies that the model predicts less relevant content in some portion of the image compared to the ground truth, and over-segmentation that the predicted image covers more pixels than the ground truth.

As observed in Tables 2 and 3, little research has until now focused on the required real-time capabilities in order to provide live feedback to clinicians during the endoscopy examinations. However, there seems to be reported systems that analyze data faster than the frame-rate threshold, and it has also been given attention in some of

the arranged competitions (see section “Competitions and Challenges”). Nonetheless, it is often a trade-off between speed (model complexity) and detection performance, indicating that this is still an important issue in future research and development of CAD systems.

Automatic Report Generation

After the endoscopist finishes an endoscopy, a high-quality report should be generated. This often a time-consuming process, where research shows that approximately one-sixth of U.S. physicians working time is spent on administrative tasks, taking time away from direct-patient care and lessening job satisfaction [115]. Moreover, there are large variations in endoscopists’ interpretations of findings as well as reporting styles. This can, and often does, lead to inconsistencies in the final decision [37]. Hence, automated report generation could both save clinical time and help standardize endoscopy reports, and recent development in natural language processing is expected to open up new possibilities in automatic report generation [86].

A method proposed by Jing et al. [51] uses neural image captioning to create reports from x-ray images. In [121], images are analyzed by a neural network, and example images of findings similar to the one at hand and attention maps are combined to reports. Most approaches focus on image analysis as a basis, and combine this with additional information [24, 33, 118]. This of course depends on access to a database containing correct information which can be used in combination with the images. A significant challenge is different reporting standards between countries or even hospitals, making it practically impossible to create a widely adoptable software.

However, for medical experts, automatic text creation might not even be the most crucial feature of such a software: A more important aspect is their ability to understand the reasoning and decision of the underlying model, enabling them to include it in their assessment. This is discussed in the next section.

Explainability

A well-known challenge associated with deep learning based CAD systems, is limited explainability due to their inherent complexity [4]. This property has caused their notoriety as black boxes whose decision-making processes are unknown, especially to end-users [35]. The need for understanding and explaining how the systems work and which roles the different data features play in the decisions, addresses different needs in the different stages of the system's development and use. The developer of the system needs to understand how data and methods are working together, as understanding and interpretability of the output helps to determine errors in the data as well as enabling targeted failure analysis. Particularly, in the context of this AIM, the medical experts require an explanation of the system's decision to assure that it concurs with the relevant medical knowledge.

Deep learning based systems, such as CNNs, have no inherent ways of providing explanations, meaning that they must either be extended to contain explanation generators, or explanations must be obtained post hoc [1, 38, 39]. A brief overview of approaches to model explanations is shown in Table 5. Models can be designed to provide justifications for their decisions as an additional task, e.g., via a text justification generator as part of the model architecture [62]. Given a model without such a design, different approaches are available: Those which explain the properties of the decision making system itself, and those which treat the system like a black box and

provide explanations based on its emergent behavior, referred to as model dependent and model agnostic approaches, respectively. One example of the former is displaying the values of the Deep Neural Network (DNN)'s internal parameters as a heat map superimposed on the classification instance [35]. Interpreted correctly, this can provide an understanding of the system's internal decision making process. Such an approach can also be extended to include information regarding the system induced decision uncertainty (meaning the part of the uncertainty not associated with the data collection and selection process), see [113].

Among the model-agnostic methods, the explanation concept LIME (*Locally Interpretable Model-agnostic Explanations*) approximates the black-box model using an interpretable model, such as a linear model, decision tree, or falling rule list [84]. This is done in the neighborhood of the instance to explain, making the resulting explanation a local one, given that it applies to a single outcome and is based on the particular instance's characteristics, as is also the case for the aforementioned model-dependent explanations.

In contrast, global explanations capture and explain the model at large, such as feature importance ranking. One class of methods capable of producing global explanations, are those based on the game-theoretic concept of Shapley values [92], which are currently enjoying a surge of interest in the statistics and machine learning literature [27, 40, 44, 66]. Shapley values are obtained by evaluating the model using all

Table 5 The different model explanation approaches regarding when they are applied: During the model development (in-model) or after the model is finished (post-

model). Explanation methods provide insight into model behavior either locally (around a particular prediction) or globally

Category		Description	Ex.
In-model		Justification text generator as part of model architecture	[62]
Post-model	Model dependent	GradCam: Display DNN activations on image	[35]
	Model agnostic	LIME: Yields a locally interpretable model approximating the full model	[84]
		SHAP: Shapley decomposition of a conditional expectation function of the full model	[68]
	Model independent	Global non-parametric Shapley decomposition	[28]

Table 6 List of GI detection, classification and segmentation challenge examples

Challenge name	URL
MICCAI 2015 Endoscopic Vision	https://polyp.grand-challenge.org/databases/
Medico 2017	http://www.multimediaeval.org/mediaeval2017/medico/
Medico 2018	http://www.multimediaeval.org/mediaeval2018/medico/
GIANA 2018	https://giana.grand-challenge.org/Home/
EAD 2019	https://ead2019.grand-challenge.org/
Biomedia 2019	https://github.com/kelkalot/biomedia-2019
Medico 2020	https://multimediaeval.github.io/editions/2020/tasks/medico/
EndoTect 2020	https://github.com/simula/icpr-endotect-2020
EDD Challenge 2020	https://edd2020.grand-challenge.org/
EndoCV 2020	https://endocv.grand-challenge.org

possible combinations of the data features. Hence, the computational complexity increases with the number of features $|f|$ as $2^{|f|}$, and the calculation involves re-training the model for each subset of features. The latter is problematic as re-training would result in different model parameters, highlighting that Shapley values are merely model agnostic, not *independent*. The widely used SHAP (*SHapley Additive exPlanations*) package [68] circumvents these challenges in different ways for various model architectures, by calculating approximate values using background samples from the data, and for deep architectures using a similar approach as the per node attribution rules from DeepLIFT [94]. The Shapley decomposition can be computed both globally and locally, and can be formulated [68] as a special case of LIME. Shapley values can also be used to obtain model-independent explanations [28].

Competitions and Challenges

There have been a series of different challenges related to automatic analysis of endoscopy data [9, 36, 79], where CNN-based approaches have been the top performing methods for the last few years. The various tasks given have been to benchmark and develop automated systems to accurately detect, localize, and segment the abnormalities inside the GI tract. These challenges targeted different tasks from detection, localization, and segmentation of GI anomalies, colorectal polyps to artifacts presence in the GI tract (see

Table 6). These regular competitions can help the research community in the field to find to find common standards for evaluating models, benchmarking state-of-the-art methods and tools, and finding new directions to bring the field forward together.

Clinical Verification and Emerging Commercial Systems

Many research groups have presented promising research results and good performance indicators, and several AI-based commercial systems have emerged, some of which are listed in Table 7. The status of these are mostly unknown, but, for example, the GI Genius system is CE marked, but still lacks US Food and Drug Administration (FDA) approval, and EndoBRAIN-EYE is approved only in Japan. For CAD systems to be deployed for real-time examinations in clinical examination rooms, or to be used for VCE data post analysis, clinical verification is strictly necessary. Still, at the time of writing, such studies are very limited. In August 2020, Repici et al. [83] presented a randomized multi-center trial, concluding that the AI-based CAD increases the adenoma detection rate (ADR), i.e., the percentage of patients with at least one histologically proven adenoma or carcinoma, demonstrating the potential of such systems. They examined 685 patients: 341 patients using the CAD system and 344 patients using only the traditional manual examination. The system achieved an ADR of 54.8%, and the control group 40.4%. This

Table 7 Emerging commercial products

Product	Vendor	Year	URL
GI Genius AI	Medtronic/Cosmo Pharma	2019	https://www.cosmopharma.com/products/gi-genius
EndoBRAIN-EYE	Cybernet	2020	https://www.cybernet.jp/english/documents/pdf/news/press/2020/20200129.pdf
CAD-Eye	Fujifilm	2020	https://www.fujifilm.eu/eu/cadeye
Ai4Gi	Ai4Gi	2016	https://ai4gi.com
UltiVision	DocBot	2018	https://www.docbot.co/gastroenterology-and-health
DISCOVERY	Pentax	2020	https://www.pentaxmedical.com/pentax/en/95/2/DISCOVERY-new
ENDO-AID	Olympus	2020	https://www.olympus.no/medical/en/Products-and-Solutions/Products/Product/ENDO-AID.html
SOMA	Augere Medical	2018	https://augere.md

demonstrates that AI-based systems can help detect adenomas, but that further improvements are required to increase detection rates, and to detect a larger number of sessile serrated lesions (at all). Considering the limitations of the study as well as the presented performance, it is clear that there are still improvements to be made, and more clinical studies are in order.

Despite significant interest from the industry, proper standards regarding evaluation methods and reproducibility are widely lacking. In addition, industry applications seem not to have focused on model explainability or model output interpretability. These are all crucial ingredients of trustworthy applications, and industry development will hopefully follow current research trends and focus more on these in the future.

Finally, when a high-performing (research) prototype has been built and tested, meeting the requirements above, it must be approved for medical use. Robust evaluation of AI based software before implementation is needed to reduce patient and health system risk, establish trust to facilitate wide-spread adoption. The common term used for such products is AI based software as a medical device (SaMD). Regulators of the SaMD applications, including the FDA in the United States, have been guided by the Global Harmonization Task Force and International Medical Device Regulators Forum (IMDRF). The IMDRF has proposed four different risk categories for SaMD each with a different set of requirements for assessing scientific and clinical validity of the technology [42]. Within gastroenterology, CADe

and CADx technologies have not yet been classified. The current FDA process for SaMD is derived from its approval process for medical devices and will be categorized into three risk categories: Classes I, II, and III (highest risk) [105]. After risk classification, premarket submission as a 510(k) pathway or de novo pathway might be relevant to GI-based AI technologies similar to Osteoidetect [107]. Moreover, given that the AI algorithms are rapidly iterative and continuously learning, it can pose a challenge to the current regulatory process. The FDA proposed a new system of regulation for AI technologies in its Digital Health Innovation Action Plan, focused on AI technologies that rely on continuous learning and adaptation [106]. Regulators around the world have also recognized the challenges involved with AI algorithms when applied to medicine and most countries have initiated efforts to develop policies tailored for SaMD. Many of them share the core principles of designation of risk, review clinical evidence to demonstrate efficacy and safety, practices to incorporate evolving AI systems.

Summary and Conclusions

In this work, we have introduced the application of automated data analysis for GI endoscopy, and presented an overview on detection and segmentation based approaches to tackle challenges like large lesion miss-rates and interobserver variability. Recent studies have shown that deep computer

vision-based approaches seem to have the potential of improving the accuracy and overall performance in GI endoscopy by providing fully automated CAD systems acting as an additional digital eye. Nevertheless, there are still several open issues and challenges which need to be addressed before automatic analyses can be usefully integrated into clinical practice. These should be regarded as issues requiring research attention in the field.

Acknowledgments This work is funded in part by the Research Council of Norway, project number 282315 (AutoCap).

References

1. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*. 2018;6:52138–60.
2. Alexandre LA, Casteleiro J, Nobreinst N. Polyp detection in endoscopic video using svms. In: Proceeding of Knowledge Discovery in Databases (PKDD). Berlin/Heidelberg: Springer; 2007. p. 358–65.
3. Ali S, Zhou F, Braden B, Bailey A, Yang S, Cheng G, Zhang P, Li X, Kayser M, Soberanis-Mukul R, Albarqouni S, Wang X, Wang C, Watanabe S, Oksuz I, Ning Q, Yang S, Khan MA, Gao X, Rittscher J. An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. *Sci Rep*. 2020;10:2748.
4. Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Precise4Q consortium: explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak*. 2020;20(1):310. <https://europemc.org/articles/PMC7706019>
5. Ameling S, Wirth S, Paulus D, Lacey G, Vilarino F. Texture-based polyp detection in colonoscopy. In: Meinzer HP, Deserno TM, Handels H, Tolxdorff T (eds) *Bildverarbeitung für die Medizin 2009. Informatik aktuell*. Springer, Berlin, Heidelberg; 2009. p. 346–50.
6. Angermann Q, Bernal J, Sánchez-Montes C, Hammami M, Fernández-Esparrach G, Dray X, Romain O, Sánchez FJ, Histace A. Towards real-time polyp detection in colonoscopy videos: adapting still frame-based methodologies for video sequences analysis. In: Cardoso MJ, Arbel T, Luo X, Wesarg S, Reichl T, González Ballester MÁ, McLeod J, Drechsler K, Peters T, Erdt M, Mori K, Linguraru MG, Uhl A, Oyarzun Laura C, Shekhar R, editors. *Computer assisted and robotic endoscopy and clinical image-based procedures*. Cham: Springer International Publishing; 2017. p. 29–41.
7. Bai W, Oktay O, Sinclair M, Suzuki H, Rajchl M, Tarroni G, Glocker B, King A, Matthews PM, Rueckert D. Semi-supervised learning for network-based cardiac MR image segmentation. In: Descoteaux M, Maier-Hein L, Franz A, Jannin P, Collins DL, Duchesne S, editors. *Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Lecture Notes in Computer Science. Springer International Publishing; 2017. p. 253–60.
8. Bell LT, Gandhi S. A comparison of computer-assisted detection (CAD) programs for the identification of colorectal polyps: performance and sensitivity analysis, current limitations and practical tips for radiologists. *Clin Radiol*. 2018;73:593.e11–8. <https://doi.org/10.1016/j.crad.2018.02.009>.
9. Bernal J, Tajkbaksh N, Snchez FJ, Matuszewski BJ, Chen H, Yu L, Angermann Q, Romain O, Rustad B, Balasingham I, Pogorelov K, Choi S, Debard Q, Maier-Hein L, Speidel S, Stoyanov D, Brandao P, Crdova H, Snchez-Montes C, Gurudu SR, Fernndez-Esparrach G, Dray X, Liang J, Histace A. Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge. *IEEE Trans Med Imaging*. 2017;36(6):1231–49.
10. Bernal J, Sánchez FJ, Fernández-Esparrach G, Gil D, Rodríguez C, Vilarino F. Wm-dova maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. *Comput Med Imaging Graph*. 2015;43:99–111.
11. Bernal J, Sánchez J, Vilarino F. Towards automatic polyp detection with a polyp appearance model. *Pattern Recogn*. 2012;45(9):3166–82.
12. Bernal JJ, Histace A, Masana M, Angermann Q, Sánchez-Montes C, Rodriguez C, Hammami M, Garcia-Rodriguez A, Córdoba H, Romain O, Fernández-Esparrach G, Dray X, Sanchez J. Polyp detection benchmark in colonoscopy videos using GTCreator: a novel fully configurable tool for easy and fast annotation of image databases. In: Proceedings of 32nd CARS conference. Berlin; 2018.
13. Billah M, Waheed S. Gastrointestinal polyp detection in endoscopic images using an improved feature extraction method. *Biomed Eng Lett*. 2018;8(1):69–75.
14. Borgli H, Thambawita V, Smedsrød PH, Hicks S, Jha D, Eskeland SL, Randel KR, Pogorelov K, Lux M, Nguyen DTD, Johansen D, Griwodz C, Stensland HK, Garcia-Ceja E, Schmidt PT, Hammer HL, Riegler MA, Halvorsen P, de Lange T. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Sci Data*. 2020;7:283. <https://doi.org/10.1038/s41597-020-00622-y>. Springer Nature
15. Boughorbel S, Jaray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews correlation coefficient metric. *PLoS One*. 2017;12(6):e0177678.
16. Bradski G. The OpenCV library. *Dr Dobb's J Softw Tools*. 2000;120:122.

17. Brandao P, Mazomenos E, Ciuti G, Caliò R, Bianchi F, Menciassi A, Dario P, Koulaouzidis A, Arezzo A, Stoyanov D. Fully convolutional neural networks for polyp segmentation in colonoscopy. In: Medical imaging 2017: computer-aided diagnosis, vol. 10134. International Society for Optics and Photonics; 2017. p. 101340F. <https://doi.org/10.1117/12.2254361>
18. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394–424.
19. Brenner H, Kloor M, Pox CP. Colorectal cancer. Lancet. 2014;383(9927):1490–502.
20. Cave DR, Hakimian S, Patel K. Current controversies concerning capsule endoscopy. Dig Dis Sci. 2019;64(11):3040–7.
21. Cho M, Kim JH, Kong HJ, Hong KS, Kim S. A novel summary report of colonoscopy: timeline visualization providing meaningful colonoscopy video information. Int J Color Dis. 2018;33(5):549–59.
22. Costamagna G, Shah SK, Riccioni ME, Foschia F, Mutignani M, Perri V, Vecchioli A, Brizi MG, Picciocchi A, Marano P. A prospective trial comparing small bowel radiographs and video capsule endoscopy for suspected small bowel disease. Gastroenterology. 2002;123(4):999–1005.
23. Cramer H. Mathematical methods of statistics. Princeton: Princeton University Press; 1946.
24. Daniels ZA, Metaxas DN. Exploiting visual and report-based information for chest x-ray analysis by jointly learning visual classifiers and topic models. In: Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI). IEEE; 2019. p. 1270–4.
25. Enns RA, Hookey L, Armstrong D, Bernstein CN, Heitman SJ, Teshima C, Leontiadis GI, Tse F, Sadowski D. Clinical practice guidelines for the use of video capsule endoscopy. Gastroenterology. 2017;152(3):497–514.
26. Fan DP, Ji GP, Zhou T, Chen G, Fu H, Shen J, Shao L. PraNet: Parallel reverse attention network for polyp segmentation. arXiv preprint arXiv:2006.11392. 2020.
27. Frye C, Rowat C, Feige I. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. 2020.
28. Fryer D, Strümke I, Nguyen H. Explaining the data or explaining a model? Shapley values that uncover non-linear dependencies. arXiv:abs/2007.06011. 2020.
29. Ghosh T, Fattah SA, Wahid KA. CHOBS: Color Histogram of Block Statistics for automatic bleeding detection in wireless capsule endoscopy video. IEEE J Transl Eng Health Med. 2018;6(May 2017):1800112.
30. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Advances in neural information processing systems. Montreal, Canada. p. 2672–2680 (y).
31. Goyal M, Yap MH, Reeves ND, Rajbhandari S, Spragg J. Fully convolutional networks for diabetic foot ulcer segmentation. In: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC); 2017. p. 618–23.
32. Guo YB, Matuszewski B. GIANA Polyp segmentation with fully convolutional dilation neural networks. In: Proceedings of International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications; 2019. p. 632–41.
33. Han Z, Wei B, Leung S, Chung J, Li S. Towards automatic report generation in spine radiology using weakly supervised framework. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2018. p. 185–93.
34. Hewett DG, Kahi CJ, Rex DK. Efficacy and effectiveness of colonoscopy: how do we bridge the gap? Gastrointest Endosc Clin. 2010;20(4):673–84.
35. Hicks S, Riegler M, Pogorelov K, Anonsen KV, de Lange T, Johansen D, Jeppsson M, Ranheim Randel K, Losada Eskeland S, Halvorsen P. Dissecting deep neural networks for better medical image classification and classification understanding. In: Proceedings of IEEE International Symposium on Computer-Based Medical Systems (CBMS); 2018. p. 363–8.
36. Hicks S, Petlund A, de Lange T, Schmidt P, Halvorsen P, Riegler M, Smedsrød P, Haugen T, Randel K, Pogorelov K, Stensland H, Dang Nguyen DT, Lux M. Acm multimedia biomedia 2019 grand challenge overview. In: Proceedings of the ACM International Conference on Multimedia (ACM MM); 2019. p. 2563–7.
37. Hicks S, Smedsrød P, Riegler M, de Lange T, Petlund A, Eskeland S, Pogorelov K, Schmidt P, Halvorsen P. Deep learning for automatic generation of endoscopy reports. Gastrointest Endosc. 2019;89: AB77.
38. Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable ai systems for the medical domain? arXiv preprint arXiv:1712.09923. 2017.
39. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Disc. 2019;9(4):e1312.
40. Huettner F, Sunder M. Axiomatic arguments for decomposing goodness of fit according to Shapley and Owen values. Electron J Stat. 2012;6:1239–50.
41. International Agency for Research on Cancer, World Health Organization: Cancer Fact Sheets. 2020. <https://gco.iarc.fr/today/fact-sheets-cancers>
42. International Medical Device Regulators Forum (IMDRF): Software as a Medical Device (SaMD): key definitions. 2013. <http://www.imdrf.org/docs/>

- imdrf/final/technical/imdrf-tech-131209-samd-key-definitions-140901.pdf
43. Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 1125–34.
 44. Israeli O. A Shapley-based decomposition of the R-square of a linear regression. *J Econ Inequal*. 2007;5:199–212.
 45. Jha D, Riegler MA, Johansen D, Halvorsen P, Johansen HD. Doubleu-net: a deep convolutional neural network for medical image segmentation. In: Proceedings of the IEEE International Symposium on Computer-Based Medical Systems (CBMS); 2020. p. 558–64.
 46. Jha D, Ali S, Emanuelsen K, Hicks SA, Thambawita V, Garcia-Ceja E, Riegler MA, de Lange T, Schmidt PT, Johansen HD, Johansen D, Halvorsen P. Kvasir-instrument: diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. 2020.
 47. Jha D, Ali S, Johansen HD, Johansen D, Rittscher J, Riegler MA, Halvorsen P. Real-time polyp detection, localisation and segmentation in colonoscopy using deep learning. arXiv preprint arXiv:2006.11392. 2020.
 48. Jha D, Hicks SA, Emanuelsen K, Johansen HD, Johansen D, de Lange T, Riegler MA, Halvorsen P. Medico multimedia task at mediaeval 2020:automatic polyp segmentation. In: Proceedings of the MediaEval 2020 Workshop; 2020.
 49. Jha D, Smedsrød PH, Riegler MA, Halvorsen P, de Lange T, Johansen D, Johansen HD. Kvasir-SEG: a segmented polyp dataset. In: Proceedings of the International Conference on Multimedia Modeling (MMM); 2020. p. 451–62.
 50. Jha D, Smedsrød PH, Riegler MA, Johansen D, De Lange T, Halvorsen P, Johansen HD. ResUNet++: an advanced architecture for medical image segmentation. In: Proceedings of International Symposium on Multimedia (ISM); 2019. p. 225–2255.
 51. Jing B, Xie P, Xing E. On the automatic generation of medical imaging reports. arXiv preprint arXiv:1711.08195. 2017.
 52. Kaminski MF, Regula J, Kraszewska E, Polkowski M, Wojciechowska U, Didkowska J, Zwierko M, Rupinski M, Nowacki MP, Butruk E. Quality indicators for colonoscopy and the risk of interval cancer. *N Engl J Med*. 2010;362(19):1795–803.
 53. Karkanis SA, Iakovidis DK, Karras DA, Maroulis DE. Detection of lesions in endoscopic video using textural descriptors on wavelet domain supported by artificial neural network architectures. In: Proceedings the International Conference on Image Processing; 2001. p. 833–6.
 54. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2019. p. 4401–10.
 55. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T. Analyzing and improving the image quality of StyleGAN. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2020.
 56. Khan A, Sohail A, Zahoor U, Qureshi AS. A survey of the recent architectures of deep convolutional neural networks. *Artif Intell Rev*. 2020;53(8):5455–516.
 57. Ko WJ, An P, Ko KH, Hahn KB, Hong SP, Cho JY. Image quality analysis of various gastrointestinal endoscopes: why image quality is a prerequisite for proper diagnostic and therapeutic endoscopy. *Clin Endosc*. 2015;48(5):374.
 58. Koulaouzidis A, Iakovidis DK, Yung DE, Rondonotti E, Kopylov U, Plevris JN, Toth E, Eliakim A, Johansson GW, Marlicz W, Mavrogenis G, Nemeth A, Thorlacius H, Tontini GE. Kid project: an internet-based digital video atlas of capsule endoscopy for research purposes. *Endosc Int Open*. 2017;5(6):E477–83.
 59. Krishnan SM, Yang X, Chan KL, Kumar S, Goh PMY. Intestinal abnormality detection from endoscopic images. In: Proceedings of the IEEE Annual International Conference of the Engineering in Medicine and Biology Society; 1998. p. 895–8.
 60. Kundu AK, Fattah SA, Rizve MN. An automatic bleeding frame and region detection scheme for wireless capsule endoscopy videos based on interplane intensity variation profile in normalized RGB color space. *J Healthc Eng*. 2018;2018:1.
 61. Le Berre C, Sandborn WJ, Aridhi S, Devignes MD, Fournier L, Smail-Tabbone M, Danese S, Peyrin-Biroulet L. Application of artificial intelligence to gastroenterology and hepatology. *Gastroenterology*. 2020;158(1):76–94.
 62. Lee H, Kim ST, Ro YM. Generation of multimodal justification using visual word constraint model for explainable computer-aided diagnosis. In: Suzuki K, Reyes M, Syeda-Mahmood T, Glocker B, Wiest R, Gur Y, Greenspan H, Madabhushi A, editors. Interpretability of machine intelligence in medical image computing and multimodal learning for clinical decision support. Cham: Springer International Publishing; 2019. p. 21–9.
 63. Lee JY, Jeong J, Song EM, Ha C, Lee HJ, Koo JE, Yang DH, Kim N, Byeon JS. Real-time detection of colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets. *Sci Rep*. 2020;10(1):1–9.
 64. Lee S, Jang B, Kim KO, Jeon S, Kwon J, Kim E, Jung J, Park K, Cho K, Kim ES, Park C, Yang C. Endoscopic experience improves interobserver agreement in the grading of esophagitis by los angeles classification: conventional endoscopy and optimal band image system. *Gut Liver*. 2014;8:154–9.
 65. Leenhardt R, Li C, Mouel JP, Rahmi G, Sabourin JC, Cholet F, Boureille A, Amiot X, Delvaux M,

- Duburque C, Leandri C, Gerard R, Leclaire S, Mesli F, Nion-Larmurier I, Romain O, Sacher-Huvelin S, Simon-Shane C, Vanbervliet G, Dray X. Cad-cap: a 25,000-image database serving the development of artificial intelligence for capsule endoscopy. *Endosc Int Open*. 2020;8:E415.
66. Lipovetsky S, Conklin M. Analysis of regression in game theory approach. *Appl Stoch Model Bus Ind*. 2001;17:319–30.
67. Liu Q, Yu L, Luo L, Dou Q, Heng PA, Heng PA. Semi-supervised medical image classification with relation-driven self-ensembling model. *IEEE Trans Med Imaging*. 2020;39:3429.
68. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. Advances in neural information processing systems, vol. 30. Curran Associates; 2017. p. 4765–74. <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>.
69. Lux M, Chatzichristofis SA. Lire: lucene image retrieval: an extensible java cbir library. In: Proceedings of the ACM International Conference on Multimedia (ACM MM); 2008. p. 10851088.
70. Madani A, Moradi M, Karargyris A, Syeda-Mahmood T. Semisupervised learning with generative adversarial networks for chest X-ray classification with ability of data domain adaptation. In: Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI); 2018. p. 1038–42.
71. Mahmud T, Paul B, Fattah SA. PolypSegNet: a modified encoder-decoder architecture for automated polyp segmentation from colonoscopy images. *Comput Biol Med*. 2020;128:104119.
72. Mirza M, Osindero S. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784. 2014.
73. Misawa M, Kudo SE, Mori Y, Hotta K, Ohtsuka K, Matsuda T, Saito S, Kudo T, Baba T, Ishida F, Itoh H, Oda M, Mori K. Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). *Gastrointest Endosc*. 2020;93(4):960–967.e3.
74. Norman B, Pedoia V, Majumdar S. Use of 2d u-net convolutional neural networks for automated cartilage and meniscus segmentation of knee mr imaging data to determine relaxometry and morphometry. *Radiology*. 2018;288(1):177–85.
75. Park SY, Sargent D, Spofford I, Vosburgh KG, A-Rahim Y. A colon video analysis framework for polyp detection. *IEEE Trans Biomed Eng*. 2012;59(5):1408–18.
76. Pogorelov K, Riegler M, Halvorsen P, Griwodz C, Lange T, Randel K, Eskeland S, Dang-Nguyen DT, Ostroukhova O, Lux M, Spampinato C. A comparison of deep learning with global features for gastrointestinal disease detection. In: CEUR Workshop Proceedings MediaEval, vol. 1984; 2017. p. 8–10.
77. Pogorelov K, Randel K, Griwodz C, de Lange T, Eskeland S, Johansen D, Spampinato C, Dang Nguyen DT, Lux M, Schmidt P, Riegler M, Halvorsen P. Kvasir: a multi-class image dataset for computer aided gastrointestinal disease detection. In: Proceedings of ACM Multimedia Systems (MMSYS); 2017.
78. Pogorelov K, Randel K, de Lange T, Eskeland S, Johansen D, Griwodz C, Spampinato C, Taschwer M, Lux M, Schmidt P, Riegler M, Halvorsen P. Nerthus: a bowel preparation quality video dataset. In: Proceedings of ACM Multimedia Systems (MMSYS); 2017.
79. Pogorelov K, Riegler M, Halvorsen P, Hicks S, Randel KR, Dang Nguyen DT, Lux M, Ostroukhova O, de Lange T. Medico multimedia task at mediaeval 2018. In: CEUR Workshop Proceedings-MediaEval; 2018.
80. Polit DF. Blinding during the analysis of research data. *Int J Nurs Stud*. 2011;48(5):636–41. <http://www.sciencedirect.com/science/article/pii/S0020748911000496>
81. Qadir HA, Balasingham I, Solhusvik J, Bergsland J, Aabakken L, Shin Y. Improving automatic polyp detection using cnn by exploiting temporal dependency in colonoscopy video. *IEEE J Biomed Health Inform*. 2020;24(1):180–93.
82. Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: understanding transfer learning for medical imaging. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS); 2019. p. 3347–57.
83. Repici A, Badalamenti M, Maselli R, Correale L, Radaelli F, Rondonotti E, Ferrara E, Spadaccini M, Alkandari A, Fugazza A, Anderloni A, Galtieri PA, Pellegatta G, Carrara S, Di Leo M, Cravotto V, Lamonaca L, Lorenzetti R, Andrealli A, Antonelli G, Wallace M, Sharma P, Rosch T, Hassan C. Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. *Gastroenterology*. 2020;159(2):512–20. <http://www.sciencedirect.com/science/article/pii/S0016508520305837>
84. Ribeiro MT, Singh S, Guestrin C. “Why should i trust you?” Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD ’16. New York: Association for Computing Machinery; 2016. p. 11351144. <https://doi.org/10.1145/2939672.2939778>.
85. Riegler M, Pogorelov K, Halvorsen P, de Lange T, Griwodz C, Schmidt PT, Eskeland SL, Johansen D. EIR Efficient computer aided diagnosis framework for gastrointestinal endoscopies. In: Proceeding of the International Workshop on Content-Based Multimedia Indexing (CBMI); 2016. p. 1–6.
86. Riegler M, Lux M, Griwodz C, Spampinato C, de Lange T, Eskeland SL, Pogorelov K, Tavanapong W, Schmidt PT, Gurrin C, Johansen D, Johansen H, Halvorsen P. Multimedia and medicine: teammates for better disease detection and survival. In: Proceedings of the ACM International Conference

- on Multimedia (ACM MM); 2016. p. 968–77. <http://doi.acm.org/10.1145/2964284.2976760>.
87. Riegler M, Pogorelov K, Eskeland SL, Schmidt PT, Albisser Z, Johansen D, Griwodz C, Halvorsen P, Lange TD. From annotation to computer-aided diagnosis: detailed evaluation of a medical multimedia system. ACM Trans Multimed Comput Commun Appl. 2017; <https://doi.org/10.1145/3079765>.
88. Rondonotti E, Soncini M, Girelli CM, Russo A, Ballardini G, Bianchi G, Cant P, Centenara L, Cesari P, Cortelezzi CC, Gozzini C, Lupinacci G, Maino M, Mandelli G, Mantovani N, Moneghini D, Morandi E, Putignano R, Schalling R, Tatarella M, Vitagliano P, Villa F, Zatelli S, Conte D, Masci E, de Franchis R. Can we improve the detection rate and interobserver agreement in capsule endoscopy? Dig Liver Dis. 2012;44(12):1006–11. <http://www.sciencedirect.com/science/article/pii/S1590865812002368>
89. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Proceeding of the International Conference on Medical image computing and computer-assisted intervention (MICCAI). Springer; 2015. p. 234–41.
90. Ross T, Reinke A, Full PM, Wagner M, Kenngott H, Apitz M, Hempe H, Filimon DM, Scholz P, Tran TN, Bruno P, Arbelez P, Bian GB, Bodenstedt S, Bolmgren JL, Bravo-Sanchez L, Chen HB, Gonzlez C, Guo D, Halvorsen P, Heng PA, Hosgor E, Hou ZG, Isensee F, Jha D, Jiang T, Jin Y, Kirtac K, Kletz S, Leger S, Li Z, Maier-Hein KH, Ni ZL, Riegler MA, Schoeffmann K, Shi R, Speidel S, Stenzel M, Twick I, Wang G, Wang J, Wang L, Wang L, Zhang Y, Zhou YJ, Zhu L, Wiesenfarth M, Kopp-Schneider A, Mller-Stich BP, Maier-Hein L. Robust medical instrument segmentation challenge 2019. 2020.
91. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432.
92. Shapley LS. A value for n-person games. Contrib Theory Games. 1953;2(28):307–17.
93. Shin Y, Qadir HA, Balasingham I. Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance. IEEE Access. 2018;6:56007–17.
94. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. CoRR abs/1704.02685. 2017. <http://arxiv.org/abs/1704.02685>
95. Silva J, Histace A, Romain O, Dray X, Granado B. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. Int J Comput Assist Radiol Surg. 2014;9(2):283–93.
96. Smetsrud PH, Gjestang HL, Nedrejord OO, Næss E, Thambawita V, Hicks SA, Borgli H, Jha D, Berstad TJD, Eskeland SL, Lux M, Espeland H, Petlund A, Dang-Nguyen DT, Garcia-Caja E, Johansen D, Schmidt PT, Hammer HL, de Lange T, Riegler M, Halvorsen P. Kvasir-capsule, a video capsule endoscopy dataset. OSF Preprints. 2020. <https://doi.org/10.31219/osf.io/gr7bn>
97. Tajbakhsh N, Gurudu SR, Liang J. Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans Med Imag. 2015;35(2):630–44.
98. Thambawita V, Jha D, Hammer HL, Johansen HD, Johansen D, Halvorsen P, Riegler MA. An extensive study on cross-dataset bias and evaluation metrics interpretation for machine learning applied to gastrointestinal tract abnormality classification. ACM Trans Comput Healthc. 2020;1(3). <https://doi.org/10.1145/3386295>
99. Thambawita V, Jha D, Riegler M, Halvorsen P, Hammer HL, Johansen H, Johansen D. The medico-task 2018: disease detection in the gastrointestinal tract using global features and deep learning. In: CEUR Workshop Proceedings -MediaEval; 2018.
100. Tomar NK, Jha D, Ali S, Johansen HD, Johansen D, Riegler MA, Halvorsen P. DDANet: Dual Decoder Attention Network forAutomatic Polyp Segmentation. arXiv preprint arXiv:2006.11392. 2020.
101. Tommasi T, Tuytelaars T. A testbed for cross-dataset analysis. In: European Conference on Computer Vision. Springer; 2014. p. 18–31.
102. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56.
103. Torralba A, Efros AA. Unbiased look at dataset bias. In: Proceedings of the International Conference on Pattern Recognition (CVPR). IEEE; 2011. p. 1521–8.
104. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. CA Cancer J Clin. 2015;65(2):87–108.
105. U.S. Food and Drug Administration: Learn if a medical device has been cleared by FDA for marketing. 2017. <https://www.fda.gov/medical-devices/consumers-medical-devices/learn-if-medical-device-has-been-cleared-fda-marketing>
106. U.S. Food and Drug Administration: Digital health innovation action plan. 2018. <https://www.fda.gov/media/106331/download>
107. U.S. Food and Drug Administration: FDA permits marketing of artificial intelligence algorithm for aiding providers in detecting wrist fractures. 2018. <https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-algorithm-aiding-providers-detecting-wrist-fractures>
108. Van Doorn SC, Hazewinkel Y, East JE, Van Leerdam ME, Rastogi A, Pellisé M, Sanduleanu-Dascalescu S, Bastiaansen BA, Fockens P, Dekker E. Polyp morphology: an interobserver evaluation for the Paris classification among international experts. Am J Gastroenterol. 2015;110(1):180.
109. Wang P, Xiao X, Brown J, Berzin T, Tu M, Xiong F, Hu X, Liu P, Song Y, Zhang D, Yang X, Li L, He J,

- Yi X, Liu J, Liu X, Lai L. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. *Nat Biotechnol*. 2018;2:741–8.
110. Wang Y, Tavanapong W, Wong J, Oh J, de Groen PC. Part-based multiderivative edge cross-sectional profiles for polyp detection in colonoscopy. *IEEE J Biomed Health Inform*. 2014;18(4):1379–89.
111. Wang Y, Tavanapong W, Wong J, Oh JH, De Groen PC. Polypalert: near real-time feedback during colonoscopy. *Comput Methods Progr Biomed*. 2015;120(3):164–79.
112. Wei J, Suriawinata A, Vaickus L, Ren B, Liu X, Lisovsky M, Tomita N, Abdollahi B, Kim A, Snover D, Baron J, Barry E, Hassanpour S. Evaluation of a deep neural network for automated classification of colorectal polyps on histopathologic slides. *JAMA Netw Open*. 2020;3:e203398.
113. Wickstrøm K, Kampffmeyer M, Jenssen R. Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. *Med Image Anal*. 2020;60:101619. <http://www.sciencedirect.com/science/article/pii/S1361841519301574>
114. Wickstrøm K, Kampffmeyer M, Jenssen R. Uncertainty modeling and interpretability in convolutional neural networks for polyp segmentation. In: Proceedings of the IEEE international workshop on machine learning for signal processing (MLSP). IEEE; 2018. p. 1–6.
115. Woolhandler S, Himmelstein DU. Administrative work consumes one-sixth of U.S. physicians working hours and lowers their career satisfaction. *Int J Health Serv*. 2014;44(4):63542.
116. Wu H, Prasad S. Semi-supervised deep learning using pseudo labels for hyperspectral image classification. *IEEE Trans Image Process*. 2018;27(3):1259–70.
117. Xie Q, Luong MT, Hovy E, Le QV. Self-training with Noisy Student improves ImageNet classification. arXiv. 2020. <http://arxiv.org/abs/1911.04252>
118. Xue Y, Xu T, Long LR, Xue Z, Antani S, Thoma GR, Huang X. Multimodal recurrent model with attention for automated radiology report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2018. p. 457–66.
119. Yang J, Faraji M, Basu A. Robust segmentation of arterial walls in intravascular ultrasound images using dual path u-net. *Ultrasonics*. 2019;96:24–33.
120. Zhang C, Tavanapong W, Wong J, de Groen PC, Oh J. Real data augmentation for medical image classification. In: Cardoso MJ, Arbel T, Lee SL, Cheplygina V, Balocco S, Mateus D, Zahnd G, Maier-Hein L, Demirci S, Granger E, Duong L, Carboneau MA, Albarqouni S, Carneiro G, editors. Intravascular imaging and computer assisted stenting, and large-scale annotation of biomedical data and expert label synthesis, vol. 10552. Springer International Publishing; 2017. p. 67–76. http://link.springer.com/10.1007/978-3-319-67534-3_8
121. Zhang Z, Xie Y, Xing F, McGough M, Yang L. Mdnet: a semantically and visually interpretable medical image diagnosis network. In: Proceedings of IEEE CVPR; 2017. p. 6428–36.
122. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. Unet ++: a nested u-net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer; 2018. p. 3–11.
123. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 2223–32.



AIM in Endoscopy Procedures

67

Aldo Marzullo, Sara Moccia, Francesco Calimeri, and
Elena De Momi

Contents

Introduction	940
Applications of Artificial Intelligence to Endoscopy Practice	942
Detection and Diagnosis During Endoscopic Procedure	942
Informative Frame Selection	943
Mosaicking and Surface Reconstruction	944
Augmented Reality Systems for Intraoperative Assistance and Surgeon Training	945
Discussion and Perspectives	946
Conclusion	947
References	947

Supplementary Information: The online version of this chapter (https://doi.org/10.1007/978-3-030-64573-1_164) contains supplementary material, which is available to authorized users.

A. Marzullo (✉) · F. Calimeri
Department of Mathematics and Computer Science,
University of Calabria, Rende, Italy
e-mail: marzullo@mat.unical.it; francesco.calimeri@unical.it; calimeri@mat.unical.it

S. Moccia
The BioRobotics Institute, Scuola Superiore Sant'Anna,
Pisa, Italy

Department of Excellence in Robotics & AI, Scuola
Superiore Sant'Anna, Pisa, Italy
e-mail: s.moccia@staff.univpm.it

E. De Momi
Department of Electronics, Information and
Bioengineering, Politecnico di Milano, Milan, Italy
e-mail: elena.demomi@polimi.it

Abstract

Artificial intelligence (AI) is revolutionizing the way medicine is practiced. In this context, the application of AI algorithms in endoscopy is gaining increasing attention so that modern endoscopy is moving towards more and more assisted/automatic solutions. Several approaches have been carried out in order to improve accuracy in diagnosis and surgical procedures. In this chapter, a general overview of the main contributions in the field is surveyed. Four main categories of applications were identified, namely, (i) detection and diagnosis during endoscopic procedure, (ii) informative frame selection, (iii) mosaicking and surface reconstruction, (iv) augmented reality systems for intraoperative assistance and surgeon training. Discussions on future

research directions and implementation in clinical practice are provided.

Keywords

Endoscopy · Artificial intelligence · Machine learning · Deep learning · Computer-aided detection · Computer-aided diagnosis · Informative-frame selection · Mosaicking · Augmented reality

Introduction

Endoscopy is a minimally invasive procedure performed to examine internal organs of the human body: a flexible tube with a light and a camera attached is inserted into the patient body to observe details of internal organs and tissues (Fig. 1). Endoscopy can also be used to carry out other tasks including minor surgery. To this aim, the endoscope can be equipped with surgical tools through the so-called *working channel*. Physical examination for diagnosing specific diseases in the small intestine can also be performed through Capsule Endoscopy (CE), a camera-embedded pill-shaped device that passes through the gastrointestinal tract, captures and transmits images to an external receiver [1]. As the capsule travels through the digestive tract, it takes thousands of pictures, which are transmitted to an external receiver.

Endoscopy is useful for investigating many anatomical districts: respiratory tract, including nose (rhinoscopy) and lower respiratory tract (bronchoscopy); the urinary tract (cystoscopy); the esophagus and the stomach (gastroscopy); the colon (colonoscopy); abdominal or pelvic cavity (laparoscopy), interior of a joint (arthroscopy), organs of the chest (thoracoscopy and mediastinoscopy) [2]. With its origin in the late 50s, modern endoscopy procedures have become the most important methods to diagnose and treat a wide variety of diseases and injuries inside the human body [3].

Endoscopy is generally considered a relatively safe procedure, and it is typically performed while the patient is conscious. However, it can be a

burden for the surgeon, whose performance could be altered by fatigue, stress, or limited experience [4]. With the rise of artificial intelligence (AI) in medicine, including recent advances in computer vision, a number of researchers have been carried out in order to improve performance in the field, so that modern endoscopy is moving toward more and more assisted/automatic procedures.

AI-assisted endoscopy is based on computer algorithms that mimics human cognitive function. The idea is based on the algorithm ability to rationalize and take actions that have the best chance of achieving a specific goal. In particular, among the variety of AI techniques which have been successfully applied in the field of endoscopy, the most significant achievements can be attributed to machine learning (ML). ML is a class of generic algorithms used to recognize patterns in data. ML algorithms learn from the experience without being explicitly programmed, improving their abilities in a trial-and-error iterative process. In this context, a significant breakthrough was Deep Learning (DL), a fast-growing subset of ML algorithms, biologically inspired by the human brain. In particular, Convolutional Neural Networks (CNNs) have been proven to be well suited for image and video processing tasks. CNNs leverage the multiple network layers (consecutive convolutional operations performed on image patches) to extract the key features from an image and provide a final classification through the fully connected layers as the output. Such a relatively simple idea has allowed to achieve impressive results in computer vision, including endoscopy-related tasks.

Several applications of AI in endoscopy can be found in literature. An overview is illustrated in Fig. 2. Among the most effective applications, detection and diagnosis tasks are one the most studied. In particular, the research interest seems to be mainly related to the gastrointestinal tract with polyp and detection and classification, lesion patterns identification, benign and precancerous polyps differentiation being the most diffused applications [5]. By naive searching the words “Artificial Intelligence,” “Machine Learning,” “Deep Learning” and “Endoscopy” on Pubmed,

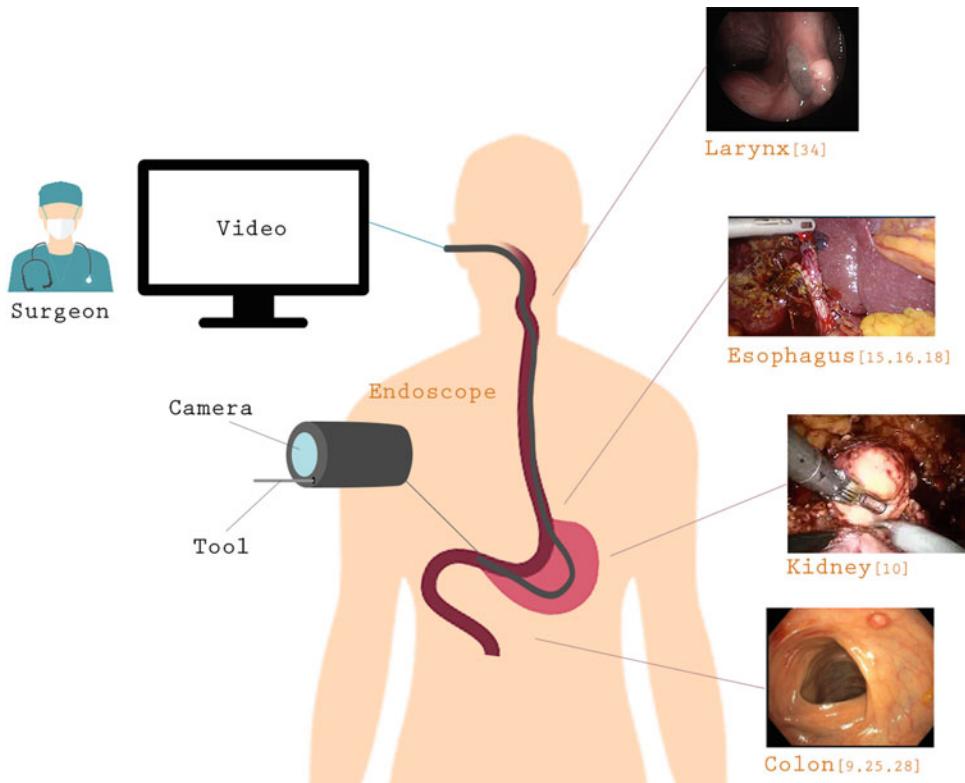


Fig. 1 Endoscopy procedure: the endoscope is inserted into the patient body and used to observe details of internal organs and tissues. It can also be used to carry out other tasks including minor surgery

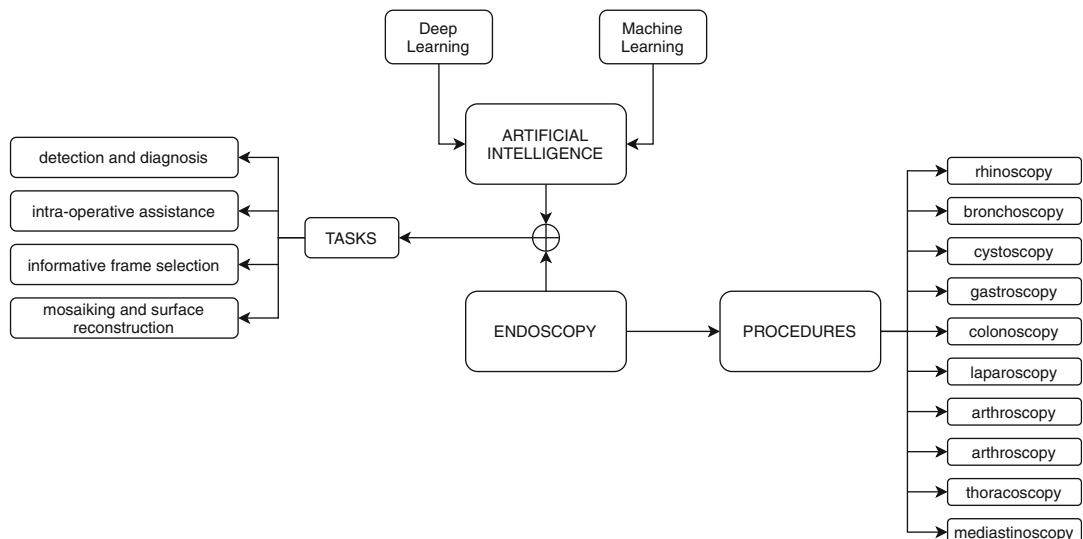


Fig. 2 Overview on applications of Artificial Intelligence in Endoscopy

indeed, about 220 results have been added in the latest 20 years, with 70 related to “diagnosis,” “detection” and “characterization” and more than 80 studies investigating the gastrointestinal tract. However, either fast-growing and milestones applications still exist. This includes real time applications of computer-aided systems, informative frame selection in video endoscopy, mosaicking and surface reconstruction.

In this chapter, those contributions will be investigated in detail. An overview of the most well studied tasks will be provided, with focus on classical methods and recent trends. Existing challenges will be analyzed to collect insight for future work.

Applications of Artificial Intelligence to Endoscopy Practice

In the following, a general overview of the main topic related to AI in Endoscopy will be provided.

Detection and Diagnosis During Endoscopic Procedure

Medical applications of AI have made significant progress during the past several years. In the field of Minimally Invasive Surgery (MIS), computer vision systems are nowadays a trendy subject. Among the possible tasks, computer-assisted detection (CADe) and computer-assisted diagnosis (CADx) systems have received particular attention from the Surgical Data Science (SDS) community [3]. CADe systems offer assistance for locating elements inside the field of view, and several techniques have been proposed to assist endoscopists to mark conspicuous structures and sections both in real-time and offline applications. Usually, this is achieved by placing a bounding box surrounding the region of interest over the image. Alternatively, a *segmentation* task can be performed where the contour of the target element is delineated [6]. Eventually, instead of just detecting the region of interest, CADx systems are used to characterize and classify elements during standard screening and

surveillance (e.g., to assess the stage of a particular form of cancer or diagnose certain infections) [4].

Several techniques have been applied from the ‘90s, initially based on image processing techniques (region growing, thresholds, etc.) and handcrafted features [7, 8], then combined with more sophisticated ML algorithms in order to improve performances [9]. However, a significant breakthrough was achieved with the rise of DL, where features can be automatically extracted from classic endoscopic images without the need of magnifying Narrow Band Imaging (NBI), Flexible spectral Imaging Color Enhancement (FICE), or Blue Laser Imaging (BLI), which are clinically useful in discriminating cancerous areas from noncancerous areas [10]. Both single frame analysis or video sequence processing approaches have been proposed. One of the most studied applications has been in computer-aided detection and computer-aided diagnosis of polyps [11]. Polyps are an abnormal growth of tissue projecting from a mucous membrane. Polyps in the colon are the most common, but it is also possible to develop polyps in other anatomical districts, including stomach and uterus. Despite most polyps being harmless (benign), some of them can eventually develop into cancer. Automatic polyp detection and classification has been an active research topic during the last 20 years and different approaches have been proposed [12, 13]. However, applications for routine patient treatment are very limited [9]. Real-time constraints, great variation of polyps appearance, the presence of other elements such as folds, blood vessels and lumen can impact on performance, and make development of CADx and CADe systems very challenging [9].

AI has also been used to develop diagnostic and detection systems associated with several diseases in several parts of the body, including the esophagus, stomach, small bowel, and colon. For example, optical diagnosis of early dysplasia relating to Barrett’s esophagus (BE) can be currently precisely done only by experts [10]. Barrett’s esophagus is a determinant condition for the development of esophageal adenocarcinoma, and automated identification of

dysplastic change in BE is one of the popular trends studying CADe gastroscopy. Nevertheless, several AI-based techniques have been developed, which achieved promising results [12, 14]. In this context, early work considered appearance features (color, vasculature and surface pattern) to train classic classifiers such as Support Vector Machines (SVM) [15]. Van der Sommen et al. [16] proposed a method based on specific texture, color filters, and machine learning for neoplasia detection. One-hundred images from 44 patients were considered in the study, achieving a sensitivity and specificity of 83%. To the same aim, Swager et al. [17] tested an AI model on 60 near microscopic resolution scans from 19 patients. Results showed a sensitivity and specificity of 90% and 93%, respectively. Recently, advanced DL techniques have been proposed with the joint of objective of increase performance and avoid the time-consuming step of manually specifying features [15]. De Groof et al. [18], proposed a DL CADe system, suitable for use in real time in clinical practice, to improve endoscopic detection of early neoplasia in patients with BE. The method was tested using 1704 unique esophageal high-resolution images derived from 669 patients, achieving more than 90% of accuracy.

Identification of esophageal squamous cell carcinoma (SCC) has also received particular attention [14]. Squamous cell carcinoma is the most prevalent esophageal cancer worldwide occurring most often in the upper and middle portions of the esophagus [19]. Guo et al. developed a real-time DL system for diagnosis of precancerous esophageal lesions [20]. The model creates a probability map indicating suspected areas of neoplasia and noncancerous areas. It was tested using more than 6000 NBI images, obtaining a sensitivity of 98.04%. Ohmori et al. used neural networks for esophageal lesions detection and differentiation [21]. More than 10,000 endoscopic nonmagnified images were used, achieving sensitivity, specificity, and accuracy were 90%, 76%, and 81%, respectively.

AI can offer valuable assistance in the management of stomach diseases. Among those applications, detection of gastric cancers and recognition

of Helicobacter pylori (HP) infection are widely studied. HP-associated chronic gastritis can cause mucosal atrophy and intestinal metaplasia, both of which increase the risk of gastric cancer [22]. In Itoh et al. [22], a CNN was used to accurate diagnosis of HP infection. The system was tested using 179 upper gastrointestinal endoscopy images obtained from 139 patients. The sensitivity and specificity of the CNN for the detection of HP infection were 86.7% and 86.7%, respectively, and the area under the ROC was 0.95.

Several studies have also demonstrated applications of AI in wireless capsule endoscopy (WCE). WCE is a technology developed for the endoscopic exploration of the small bowel [23, 24], with cases for which WCE proved useful increasing significantly over the last few years. A major challenge of WCE for clinicians is the time-intensive nature of reviewing the images [25]. For this reason, ML, and in particular DL, represents a remarkable perspective and contributes to broadening the range of applicability. According to recent related reviews, indeed, ML has been used to enhance WCE analysis in a wide range of studies, including detection of GI bleeding, detection of small intestinal ulcer and erosion, detection of GI angioectasia, among others [23].

Finally, CAD systems, in particular CADe, are not only limited to disease detection and characterization. Considering the number of elements which can be present in a traditional endoscopy procedure, several researchers started to develop systems to detect and track other objects outside anatomical tracts [26]. Instrument detection, segmentation, tracking and pose estimation are some examples in this context. However, illumination levels, variations in background and the different number of tools in the field of view, pose difficulties to algorithm and model training [27].

Informative Frame Selection

AI is aiding in improving or monitoring endoscopic quality. In this context, informative frame selection systems have been implemented to

monitor the quality standards and reducing the amount of information to process, for example, indicating degrading factors such as noise, acquisition errors, glare, blur, and uneven illumination which can lead to diagnostic errors [28]. Manual revision of endoscopic videos, indeed, is costly and time-consuming. In WCE, for example, a typical recording is composed by more than 50,000 frames and about 45–90 min are required for reviewing it [29]. To minimize these amount of data, a possible step is to eliminate the frames with no diagnostically relevant information, which only serve to degrade the accuracy of video analysis [30]. Such solutions are already having a larger impact on clinical outcomes and are therefore likely to increasingly become a field of interest in endoscopy [31].

In Alizadeh et al. [32], a multi-stage method including active contour, color range ratio, adaptive gamma correction method, canny color edge detection operator, and morphological processing [33] was proposed to discriminate between internal mucosa (informative) and uninformative regions (sensitivity and specificity of 81% and 92%). In Moccia et al. [34], a set of intensity, keypoint-based and textural features, and multi-class SVMs are used to classify informative and three classes of uninformative frames in laryngoscopic videos in NBI (mean AUC of 91% tested on a balanced set of 720 images from 18 different laryngoscopic videos).

Recently, it has been shown that deep-learning algorithms may outperform standard learning approaches for image analysis. Patrini et al. [35] presented a novel approach toward the selection of informative frames in laryngoscopic, also

showing the potential of *transfer learning* in the field. SVMs and CNN-based approaches were used to classify frames as informative and uninformative ones such as blurred, with saliva or specular reflections, and underexposed. The system was tested on a dataset of 18 NBI endoscopic videos, referring to 18 different patients affected by squamous cell carcinoma, achieving Area under the ROC curve of 0.98.

Mosaicking and Surface Reconstruction

Endoscopic procedures have the advantage of providing a minimally invasive intervention implying less burden for both patients and surgeons in order to guide the endoscope and be aware of its position inside the body. However, for the latter, such techniques require a high degree of coordination and fine motor abilities to guide the endoscope and be aware of its position inside the body. Among the causes, a very limited field of view provided by the endoscope and the absence of a direct relation between the displayed image and the physical environment [2]. These problems have raised the interest of researchers and industries in providing methods for assisting the endoscopist by creating an enhanced field of view.

Mosaicking (also called stitching) refer to the process of combining several (partially overlapping) frames from a video sequence of endoscopic images to create a broader field of view or panorama image with a wider perspective. In a typical pipeline (Fig. 3), images are acquired during the endoscopic procedure and pre-processed in order to clean and reduce the amount of input

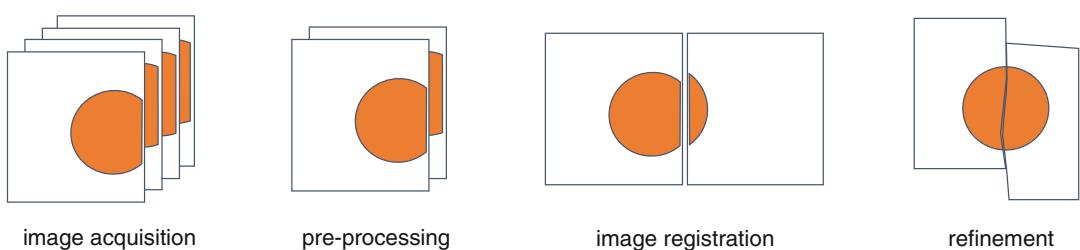


Fig. 3 Typical workflow in image mosaicking

data (e.g., subsampling the number of subsequent images in the video stream which may contain redundant information) [36, 37]. Then, image registration is performed, which is the task of aligning images representing physically connected components [2]. Common approaches include pixel-based or feature-based alignment methods. The first, consider the entire pixel group within the overlapping regions, while the latter extract higher-level features and uses their similarity to compute the match [38, 39]. Often, a further optimization process is performed on the obtained reconstruction to improve the accuracy of the final result [2]. Several techniques, originally developed by the photography and photogrammetry community, have been applied for each step. A comprehensive survey is provided in Szeliski [40], Bergen and Wittenberg [2].

Mosaicking can be applied either by projecting the images onto planar surfaces or, if the scene observed in the images has a significant 3D structure and the camera performs any translational motion, onto an higher dimensional space. Here, mosaicking and 3D surface reconstruction intersect and can be regarded as involving similar problems. In this context, existing works have proposed a software solution to reconstruct the 3D structure of a target organ (e.g., colon, liver, and larynx) with the estimated endoscope poses from an endoscope video. The methods are ranging from shape-from-shading (SfS), structure-from-motion (SfM), visual simultaneous localization and mapping (SLAM) [41, 42]. SfS attempts to retrieve the 3D structure of the human tissue surface by considering variations in illuminations and surface reflectance. Despite its simple assumptions (e.g., all visible surface points will receive the same amount of light from the same direction [43]), this method has provided promising results in endoscopy applications [44, 45]. However, SfS techniques alone may not be appropriate for scenes in which the illumination conditions drastically change during the operation. For this reason, researchers have started to combine SfS methods with SfM methods to simultaneously exploit shading and feature information for the reconstruction of surfaces from endoscopic images [46] and with

SLAM, to also exploit the endoscope trajectory in real-time [47, 48]. Bergen et al. [2] provided a detailed overview of the milestone techniques used for mosaicking and surface reconstruction, surveying relevant applications based on medical branches and anatomical structures.

Recently, DL techniques have been started to catch the interest of the researcher community. Turan et al. uses DL to 3D surface reconstruction methods for endoscopic capsule robots [49]. A particular interest has been observed for the task of predicting depth information in endoscopy from 2D images, where several deep learning approaches have been proposed with promising results [50–52]. However, several work has been conducted concerning homography estimation [53, 54] and fetoscopic mosaicking [55].

Augmented Reality Systems for Intraoperative Assistance and Surgeon Training

Endoscopic simulators are used to virtually reproduce real-life clinical settings. This is particularly important, since both physical and mental burdens for the surgeon may be reduced in the operating room, and trainees may acquire technical proficiency by practicing directly on a live patient [56]. Several possibilities are offered by Augmented Reality (AR), Virtual Reality (VR) systems: from one hand, surgeons can exploit enhanced view during surgery. On the other hand, they allow trainees to experience simulated clinical scenarios by practicing on virtual patients, test their response and benefit from CAD systems in real-time [57]. In AR the users real world is *augmented* by combining live video input with immersive display technology (e.g., computer-generated 3D objects superimposed on the video frames). The objective is to provide the surgeon with additional information about the patient while looking onto the patient. This differs from VR, where the basic elements of the environment are entirely simulated by a computer in an effort to simulate their existence [58]. Despite its long history, which can be dated back in 60s, its only in recent years that AR has started to be used

for medical practice. Localization of specific anatomical structures within the human body was one of the first applications [58, 59]. Bertrand et al. [60] proposed a technique which overlays a deformable preoperative model semi-automatically onto a laparoscopic image using a new software called Hepataug. Their analysis reported the feasibility and the potential interest of using Hepataug to achieve AR with a deformable model in laparoscopic liver resection. Hussain et al. [61] developed and assessed the performance of a video-based augmented reality system, combining preoperative computed tomography and real-time microscopic video. Virtual endoscopy image was registered to the microscope-based video of the intact tympanic membrane based on fiducial markers and a homography transformation was applied during microscope movements. In particular, the system provided additional visual information on the middle ear structures and the surgical instrument with submillimetric precision, compatible for middle ear surgery.

However, since the introduction of wearable headup displays, there has been much interest in the surgical community adapting this technology into routine surgical practice. In a recent survey by Yoon et al. [62], among 74 published articles that evaluated the utility of wearable headup displays in surgical settings, many of them were related to endoscopic surgical procedures. Recently, in Al Janabi et al. [63] the Microsoft HoloLens has been evaluated as a feasible alternative to conventional endoscopic monitors, showing that the device facilitated improved outcomes of performance and was widely accepted as a surgical visual aid by the study participants. In Qian et al. [64], the authors introduce ARsist, an augmented reality application based on an optical see-through head-mounted display, to help the first assistant perform tasks his/her task more efficiently, and hence improve the outcome of robot-assisted laparoscopic surgeries.

Discussion and Perspectives

The use of AI in endoscopy is gaining growing interest because it has the potential to increase the quality of endoscopy at many levels. Real-time and offline detection and diagnosis during

endoscopic procedure have achieved promising results. However, various issues and challenges still emerge before their effective use in clinical applications and medical routines. The first major limitation relies heavily on algorithm generalization capability. As already pointed-out in recent related work [11, 12, 23], multicenter studies should be performed in order to avoid any bias related to the machine used to acquire data. Furthermore, validation should be also focused on nontrivial instances (e.g., high quality frames or sequences, with low variability) to not overrate performances. In this context, limitations of the study should be mentioned in papers, and suggestions for an effective use in clinical settings should be included to help understanding future steps [31]. Also, interpretability of the model (especially DL) should be improved in order to increase algorithm reliability. Nevertheless, some AI products already exist that have obtained regulatory approval, and could be available for clinical use in the next future. For example, *EndoBRAIN* (Olympus, Tokyo, Japan), an AI system to support the diagnosis of colorectal polyps obtained regulatory approval and is now delivered in combination with the endocytoscopy system, therefore facilitating realtime *in vivo* characterization of the tissue [65]. A deep learning solution (GI Genius; Medtronic, Minneapolis, Minn, USA) for the improvement of polyp detection rate on ileocolonoscopy has been released [25]. CAD systems, thus, could represent a significant support for the next future of endoscopy. Novel and promising tools are about to be developed. However, careful guidelines should be provided to endoscopists to help them understand the potential of such novel technologies [31, 66, 67]. This will be crucial to determine the future of detection and diagnosis during endoscopic procedures. Furthermore, AR in endoscopy surgical practice is a valuable field of research which may significantly improve the quality of operations in the next future. The use of wearable headup displays are worth investigating, since they usually result in improved ergonomics compared to traditional procedures in which the surgeon is required to turn his or her head repeatedly to shift visual focus between the surgical field and the imaging monitor [62]. In this direction, reducing the size of

the screen as well as holographic projections represents areas of improvement.

Concerning mosaicking and surface reconstruction, several challenges can still be observed. First, improving algorithm processing time for real-time applications represents an important objective. Some applications have still reached valuable results in providing a panoramic view to be used in clinical settings, while acceleration up to real-time capability still represents an issue [2]. Finally, few work has been successfully conducted on reconstruction and mosaicking of nonrigid structures [68, 69] which achieved valuable results. However, room for improvements and novel ideas should be considered for this task which can be seen as a challenge that has not been solved yet [41]. A framework for dealing with respiration deformation, cardiac motion, organ shift and tissue tool interaction is needed. To this aim, DL solutions could be considered a valuable field to explore where models can be trained at simulating the deformation model [70, 71].

Conclusion

In conclusion, this chapter outlined the current status of research and development and the prospects for application of AI in endoscopy. Such techniques have opened up new possibilities and may become effective in clinical practice by helping physicians, but not by reducing their relevance.

References

- Rahim T, Usman MA, Shin SY. A survey on contemporary computer-aided tumor, polyp, and ulcer detection methods in wireless capsule endoscopy imaging. *Comput Med Imaging Graph.* 2020;85:101767.
- Bergen T, Wittenberg T. Stitching and surface reconstruction from endoscopic image sequences: a review of applications and methods. *IEEE J Biomed Health Inform.* 2014;20(1):304–21.
- Sharma P, Pante A, Gross SA. Artificial intelligence in endoscopy. *Gastrointest Endosc.* 2020;91(4):925–31.
- El Hajjar A, Rey JF. Artificial intelligence in gastrointestinal endoscopy: general overview. *Chin Med J.* 2020;133(3):326.
- Togashi K. Applications of artificial intelligence to endoscopy practice: the view from Japan Digestive Disease Week 2018. *Dig Endosc.* 2019;31(3):270–2.
- Dougherty G. Medical image processing: techniques and applications. Springer Science & Business Media; 2011.
- Karkanis SA, Iakovidis DK, Maroulis DE, Karras DA, Tzivras M. Computer-aided tumor detection in endoscopic video using color wavelet features. *IEEE Trans Inf Technol Biomed.* 2003;7(3):141–52.
- Iakovidis DK, Maroulis DE, Karkanis SA. An intelligent system for automatic detection of gastrointestinal adenomas in video endoscopy. *Comput Biol Med.* 2006;36(10):1084–103.
- Bernal J, Tajkabkash N, Sánchez FJ, Matuszewski BJ, Chen H, Yu L, et al. Comparative validation of polyp detection methods in video colonoscopy: results from the MICCAI 2015 endoscopic vision challenge. *IEEE Trans Med Imaging.* 2017;36(6):1231–49.
- Mori Y, Kudo SE, Mohamed HE, Misawa M, Ogata N, Itoh H, et al. Artificial intelligence and upper gastrointestinal endoscopy: current status and future perspective. *Dig Endosc.* 2019;31(4):378–88.
- Min JK, Kwak MS, Cha JM. Overview of deep learning in gastrointestinal endoscopy. *Gut Liver.* 2019;13(4):388.
- Alagappan M, Brown JRG, Mori Y, Berzin TM. Artificial intelligence in gastrointestinal endoscopy: the future is almost here. *World J Gastrointest Endosc.* 2018;10(10):239.
- Gulati S, Emmanuel A, Patel M, Williams S, Haji A, Hayee B, et al. Artificial intelligence in luminal endoscopy. *Therapeut Adv Gastrointest Endos.* 2020;13:2631774520935220.
- Lazăr DC, Avram MF, Faur AC, Goldiș A, Romoșan I, Tăban S, et al. The impact of artificial intelligence in the endoscopic assessment of premalignant and malignant esophageal lesions: present and future. *Medicina.* 2020;56(7):364.
- de Souza Jr LA, Palm C, Mendel R, Hook C, Ebigbo A, Probst A, et al. A survey on Barrett's esophagus analysis using machine learning. *Comput Biol Med.* 2018;96:203–13.
- van der Sommen F, Zinger S, Curvers WL, Bisschops R, Pech O, Weusten BL, et al. Computer-aided detection of early neoplastic lesions in Barrett's esophagus. *Endoscopy.* 2016;48(07):617–24.
- Swager AF, van der Sommen F, Klomp SR, Zinger S, Meijer SL, Schoon EJ, et al. Computer-aided detection of early Barrett's neoplasia using volumetric laser endomicroscopy. *Gastrointest Endosc.* 2017;86(5):839–46.
- de Groot AJ, Struyvenberg MR, van der Putten J, van der Sommen F, Fockens KN, Curvers WL, et al. Deep-learning system detects neoplasia in patients with Barrett's esophagus with higher accuracy than endoscopists in a multistep training and validation study with benchmarking. *Gastroenterology.* 2020;158(4):915–29.
- Marks R. Squamous cell carcinoma. *Lancet.* 1996;347(9003):735–8.
- Guo L, Xiao X, Wu C, Zeng X, Zhang Y, Du J, et al. Real-time automated diagnosis of precancerous lesions and early esophageal squamous cell carcinoma using a

- deep learning model (with videos). *Gastrointest Endosc.* 2020;91(1):41–51.
- 21. Ohmori M, Ishihara R, Aoyama K, Nakagawa K, Iwagami H, Matsuura N, et al. Endoscopic detection and differentiation of esophageal lesions using a deep neural network. *Gastrointest Endosc.* 2020;91(2):301–9.
 - 22. Itoh T, Kawahira H, Nakashima H, Yata N. Deep learning analyzes *Helicobacter pylori* infection by upper gastrointestinal endoscopy images. *Endoscopy Int Open.* 2018;6(2):E139.
 - 23. Redondo-Cerezo E, Sánchez-Capilla AD, De La Torre-Rubio P, De Teresa J. Wireless capsule endoscopy: perspectives beyond gastrointestinal bleeding. *World J Gastroenterol: WJG.* 2014;20(42):15664.
 - 24. Ashour AS, Dey N, Mohamed WS, Tromp JG, Sherratt RS, Shi F, et al. Colored video analysis in wireless capsule endoscopy: a survey of state-of-the-art. *Curr Med Imaging.* 2020;16:1074.
 - 25. Soffer S, Klang E, Shimon O, Nachmias N, Eliakim R, Ben-Horin S, et al. Deep learning for wireless capsule endoscopy: a systematic review and meta-analysis. *Gastrointest Endosc.* 2020;92:831.
 - 26. Liu Y, Zhao Z. Review of research on detection and tracking of minimally invasive surgical tools based on deep learning. *J Biomed Eng.* 2019;36(5):870–8.
 - 27. Colleoni E, Moccia S, Du X, De Momi E, Stoyanov D. Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers. *IEEE Robot Automat Lett.* 2019;4(3):2714–21.
 - 28. Park SY, Sargent D, Spofford I, Vosburgh KG, Yousif A, et al. A colon video analysis framework for polyp detection. *IEEE Trans Biomed Eng.* 2012;59(5):1408–18.
 - 29. Iddan G, Meron G, Glukhovsky A, Swain P. Wireless capsule endoscopy. *Nature.* 2000;405(6785):417.
 - 30. Fan Y, Meng MQH, Li B. A novel method for informative frame selection in wireless capsule endoscopy video. In: 2011 Annual international conference of the IEEE engineering in medicine and biology society. IEEE; 2011. p. 4864–7.
 - 31. van der Sommen F, de Groot J, Struyvenberg M, van der Putten J, Boers T, Fockens K, et al. Machine learning in GI endoscopy: practical guidance in how to interpret a novel field. *Gut.* 2020;69:2035.
 - 32. Alizadeh M, Sharzehi K, Talebpour A, Soltanian-Zadeh H, Eskandari H, Maghsoudi OH. Detection of uninformative regions in wireless capsule endoscopy images. In: 2015 41st annual northeast biomedical engineering conference. IEEE; 2015. p. 1–2.
 - 33. Szeliski R. Computer vision: algorithms and applications. Springer Science & Business Media; 2010.
 - 34. Moccia S, Vanone GO, De Momi E, Laborai A, Guastini L, Peretti G, et al. Learning-based classification of informative laryngoscopic frames. *Comput Methods Prog Biomed.* 2018;158:21–30.
 - 35. Patrini I, Ruperti M, Moccia S, Mattos LS, Frontoni E, De Momi E. Transfer learning for informative-frame selection in laryngoscopic videos through learned features. *Med Biol Eng Comput.* 2020;58:1225–38.
 - 36. Miranda-Luna R, Hernandez-Mier Y, Daul C, Blondel WC, Wolf D. Mosaicing of medical video-endoscopic images: data quality improvement and algorithm testing. In: (ICEEE). 1st international conference on electrical and electronics engineering, 2004. IEEE; 2004. p. 530–5.
 - 37. Weibel T, Daul C, Wolf D, Rösch R, Guillemin F. Graph based construction of textured large field of view mosaics for bladder cancer diagnosis. *Pattern Recogn.* 2012;45(12):4138–50.
 - 38. Behrens A, Stehle T, Gross S, Aach T. Local and global panoramic imaging for fluorescence bladder endoscopy. In: 2009 Annual international conference of the IEEE engineering in medicine and biology society. IEEE; 2009. p. 6990–3.
 - 39. Iakovidis DK, Spyrou E, Diamantis D. Efficient homography-based video visualization for wireless capsule endoscopy. In: 13th IEEE international conference on bioinformatics and bioengineering. IEEE; 2013. p. 1–4.
 - 40. Szeliski R. Image alignment and stitching: a tutorial. *Found Trends® Comput Graph Vis.* 2006;2(1):1–104.
 - 41. Maier-Hein L, Mountney P, Bartoli A, Elhawary H, Elson D, Groch A, et al. Optical techniques for 3D surface reconstruction in computer-assisted laparoscopic surgery. *Med Image Anal.* 2013;17(8):974–96.
 - 42. Münzer B, Schoeffmann K, Bösörmenyi L. Content-based processing and analysis of endoscopic images and videos: a survey. *Multimed Tools Appl.* 2018;77(1):1323–62.
 - 43. Prados E, Faugeras O. Shape from shading. In: Handbook of mathematical models in computer vision. Springer; 2006. p. 375–88.
 - 44. Wang R, Price T, Zhao Q, Frahm JM, Rosenman J, Pizer S. Improving 3D surface reconstruction from endoscopic video via fusion and refined reflectance modeling. In: Medical imaging 2017: image processing, vol. 10133. International Society for Optics and Photonics; 2017. p. 101330B.
 - 45. Zhao Q, Price T, Pizer S, Niethammer M, Alterovitz R, Rosenman J. The endoscopogram: a 3D model reconstructed from endoscopic video frames. In: International conference on medical image computing and computer-assisted intervention. Springer; 2016. p. 439–47.
 - 46. Phan TB, Trinh DH, Wolf D, Daul C. Optical flow-based structure-from-motion for the reconstruction of epithelial surfaces. *Pattern Recogn.* 2020;105:107391.
 - 47. Park J, Hwang Y, Yoon JH, Park MG, Kim J, Lim YJ, et al. Recent development of computer vision technology to improve capsule endoscopy. *Clin Endosc.* 2019;52(4):328.
 - 48. Qiu L, Ren H. Endoscope navigation with SLAM-based registration to computed tomography for transoral surgery. *Int J Intell Robot Appl.* 2020;4(2):252–63.

49. Turan M, Pilavci YY, Ganiyusufoglu I, Araujo H, Konukoglu E, Sitti M. Sparse-then-dense alignment-based 3D map reconstruction method for endoscopic capsule robots. *Mach Vis Appl.* 2018;29(2):345–59.
50. Yoon JH, Park MG, Hwang Y, Yoon KJ. Learning depth from endoscopic images. In: 2019 International conference on 3D vision. IEEE; 2019. p. 126–34.
51. Mahmood F, Durr NJ. Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy. *Med Image Anal.* 2018;48:230–43.
52. Rau A, Edwards PE, Ahmad OF, Riordan P, Janatka M, Lovat LB, et al. Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. *Int J Comput Assist Radiol Surg.* 2019;14(7):1167–76.
53. DeTone D, Malisiewicz T, Rabinovich A. Deep image homography estimation. arXiv preprint arXiv:160603798. 2016.
54. Gomes S, Valério MT, Salgado M, Oliveira HP, Cunha A. Unsupervised neural network for homography estimation in capsule endoscopy frames. *Procedia Comput Sci.* 2019;164:602–9.
55. Bano S, Vasconcelos F, Tella-Amo M, Dwyer G, Gruijthuijsen C, Vander Poorten E, et al. Deep learning-based fetoscopic mosaicking for field-of-view expansion. *Int J Comput Assist Radiol Surg.* 2020;15:1807–16.
56. Tokuyasu T, Okamura W, Kusano T, Inomata M, Shiraishi N, Kitanou S. Training system for endoscopic surgery by using augmented reality and forceps control devices. In: 2014 Ninth international conference on broadband and wireless computing, communication and applications. IEEE; 2014. p. 541–4.
57. Bhushan S, Anandasabapathy S, Shukla R. Use of augmented reality and virtual reality technologies in endoscopic training. *Clin Gastroenterol Hepatol.* 2018;16(11):1688–91.
58. Mahmud N, Cohen J, Tsourides K, Berzin TM. Computer vision and augmented reality in gastrointestinal endoscopy. *Gastroenterol Report.* 2015;3(3):179–84.
59. Shuhaiber JH. Augmented reality in surgery. *Arch Surg.* 2004;139(2):170–4.
60. Bertrand LR, Abdallah M, Espinel Y, Calvet L, Pereira B, Ozgur E, et al. A case series study of augmented reality in laparoscopic liver resection with a deformable preoperative model. *Surg Endosc.* 2020;34:5642. <https://doi.org/10.1007/s00464-020-07815-x>.
61. Hussain R, Lalande A, Marroquin R, Guigou C, Grayeli AB. Video-based augmented reality combining CT-scan and instrument position data to microscope view in middle ear surgery. *Sci Rep.* 2020;10(1):1–11.
62. Yoon JW, Chen RE, Kim EJ, Akinduro OO, Kerezoudis P, Han PK, et al. Augmented reality for the surgeon: systematic review. *Int J Med Robot Comput Assist Surg.* 2018;14(4):e1914.
63. Al Janabi HF, Aydin A, Palaneer S, Macchione N, Al-Jabir A, Khan MS, et al. Effectiveness of the HoloLens mixed-reality headset in minimally invasive surgery: a simulation-based feasibility study. *Surg Endosc.* 2020;34(3):1143–9.
64. Qian L, Deguet A, Kazanzides P. ARsist: augmented reality on a head-mounted display for the first assistant in robotic surgery. *Healthc Technol Lett.* 2018;5(5):194–200.
65. Neumann H, Bisschops R. Artificial intelligence and the future of endoscopy. *Dig Endosc.* 2019;31(4):389–90.
66. Namikawa K, Hirasawa T, Yoshio T, Fujisaki J, Ozawa T, Ishihara S, et al. Utilizing artificial intelligence in endoscopy: a clinicians guide. *Expert Rev Gastroenterol Hepatol.* 2020;14:689.
67. Bilal M, Brown JRG, Berzin TM. Incorporating standardised reporting guidelines in clinical trials of artificial intelligence in gastrointestinal endoscopy. *Lancet Gastroenterol Hepatol.* 2020;5:962.
68. Li Y, Richter F, Lu J, Funk EK, Orosco RK, Zhu J, et al. SuPer: a surgical perception framework for endoscopic tissue manipulation with surgical robotics. *IEEE Robot Automat Lett.* 2020;5(2):2294–301.
69. Turan M, Almaliooglu Y, Araujo H, Konukoglu E, Sitti M. A non-rigid map fusion-based direct SLAM method for endoscopic capsule robots. *Int J Intell Robot Appl.* 2017;1(4):399–409.
70. Shimada S, Golyanik V, Theobalt C, Stricker D. IsMoGAN: adversarial learning for monocular non-rigid 3D reconstruction. In: IEEE conference on computer vision and pattern recognition workshops. 2019. p. 0–0.
71. Bozic A, Zollhofer M, Theobalt C, Nießner M. Deepdeform: learning non-rigid RGB-D reconstruction with semi-supervised data. In: IEEE/CVF conference on computer vision and pattern recognition. 2020. p. 7002–12.



AIM in Barrett's Esophagus

68

Joost van der Putten and Fons van der Sommen

Contents

Introduction	952
Problem Statement	952
The Case for AI in the Esophagus	952
AI for Barrett's Esophagus	953
BE Cancer Detection Using White Light Endoscopy	953
BE Cancer Detection Using Narrow-Band Imaging	957
BE Cancer Detection Using Endomicroscopy	958
AI for Quality Assessment in the Esophagus	959
Discussion	960
References	961

Abstract

The imaging quality in gastroenterology has increased dramatically in the last decade. Current state-of-the-art endoscopy devices are capable of capturing high-definition footage with white light endoscopy (WLE) and several optical chromoscopy techniques. This strong improvement has caused a paradigm shift in the field of endoscopy from visualization to interpretation. At the same time, the field of

computer science has developed at a rapid pace to a point where current computational hardware can be employed to develop sophisticated artificial intelligence (AI) systems for a huge variety of applications, including endoscopic imaging. Prognosis of Barrett's cancer depends largely on the stage at the time of diagnosis. Unfortunately, the majority of cancers are only found at an advanced stage. Since most endoscopies are performed at peripheral hospitals, where specific expertise on – and familiarity with – the visual representation.

Keywords

Barrett's esophagus · Artificial intelligence · Deep learning

J. van der Putten (✉) · F. van der Sommen
Eindhoven University of technology, VCA group,
Eindhoven, The Netherlands
e-mail: j.a.v.d.putten@tue.nl

Introduction

Problem Statement

Esophageal cancer is the eighth most common cancer and the sixth leading cause of cancer-related death worldwide [1]. Despite vast improvements in image quality of endoscopes over the years, detection of early lesions fails frequently when surveillance is carried out by general endoscopists who must undertake the bulk of surveillance [2]. Additionally, 80% of all cases occur in less-developed regions, which have less access to experienced physicians as well as specialized equipment. These factors highlight the need for a reliable, cheap, and universal solution for the detection of cancer in the esophagus.

The majority of esophageal cancers can be subdivided into two main histological subtypes: esophageal adenocarcinomas (EAC) and esophageal squamous cell carcinomas (ESCC). A pre-existing condition called Barrett's esophagus (BE) is typically the origin of EACs. BE develops in the lower third of the esophagus when the normal mucosal cells lining the lower portion of the esophagus are replaced with acid-resistant cells that are normally only present in the small intestine and large intestine. In contrast, SCC occurs mostly in flat cells lining the upper two-thirds of the esophagus [3]. Unfortunately, the prognosis for both EACs and ESCC is poor, since the lesions are predominantly identified at an advanced stage. Early detection is paramount as randomized trials have demonstrated that low-morbidity endoscopic treatment (e.g., endoscopic eradication of dysplasia via ablation or resection) reduces the risk of developing late-stage cancer significantly [4, 5]. Endoscopic resection of esophageal cancer at an early stage (i.e., limited to the mucosa or muscularis mucosa) can achieve >90% cure rates without the need for esophagectomy, which carries substantial morbidity [6, 7].

The Case for AI in the Esophagus

The quality of imaging in gastroenterology has increased dramatically in the last few decades.

Current state-of-the-art endoscopy devices are capable of capturing high-definition footage with white light endoscopy (WLE) and several optical chromoscopy techniques. This has caused a paradigm shift in the field of endoscopy from visualization to emphinterpretation. At the same time, the field of computer science has developed at a rapid pace to a point where current computational hardware can be employed to develop sophisticated AI systems for a huge variety of applications [8, 9], including endoscopic imaging [10, 11]. These two factors combined have paved the way for effective use of artificial intelligence in the esophagus. Especially advances in the field of deep learning have had a tremendous impact in the number of such systems [12].

Properly designed and tested AI systems are extremely valuable for the field of endoscopy. The most obvious use case is the early detection of cancer. Both EAC and SCC are especially deadly diseases, since many early lesions are missed, which then progress to cancers for which endoscopic treatment is no longer possible [13]. Despite advances in therapy, the prognosis for patients diagnosed with esophageal cancer remains poor with population-based studies showing a 5-year survival rate of <20% [14, 15]. Prognosis of esophageal cancer depends largely on the stage at the time of diagnosis. Unfortunately, the majority of cancers are only found at an advanced stage [16, 17]. Since most endoscopies are performed at peripheral hospitals, where specific expertise on – and familiarity with – the visual representation of early lesions is typically not present, a computer-aided detection system could save countless lives when performance of the AI is comparable to experts.

Visual identification of early cancer in the esophagus is extremely difficult and a highly subjective task that requires significant effort and experience. Even experts score a sensitivity of only 48% using the most common WLE modality [18]. For this reason, the Seattle protocol was introduced [19], in which biopsies are taken every 1–2 cm in four quadrants. In practice, however, adherence to the protocol is low and research has found that endoscopists systematically undersample patients with long-segment BE

[20]. This leads to more missed lesions and progression to cancer. A biopsy-indicator algorithm could alleviate these problems significantly.

The effectiveness of AI systems is highly dependent on the quality of the endoscopic inspection [21]. A given computer-aided system with perfect accuracy theoretically solves many problems. However, if a lesion is not properly imaged, such systems cannot be used to their full potential. For this reason, patients can greatly benefit from automatic quality inspection of surveillance endoscopies as well. Patients are generally only partially sedated so the esophagus is not impeded from moving during routine endoscopies, especially when inexperienced doctors perform the operation, and patient reflexes cause contractions which may lead to missed lesions. AI can help physicians in this context by either providing feedback about video quality or by automatically selecting the best frames for further analysis.

While artificial intelligence is on the rise in many research and clinical areas, including gastrointestinal imaging, many challenges remain. Simultaneous integration of the technical and medical aspects of these types of studies is often difficult and physical implementation of these systems is much more strenuous than in other fields that are not bound by privacy laws [22]. This work attempts to showcase the research that has been done pertaining to AI in Barrett's esophagus which will help with bridging this gap.

AI for Barrett's Esophagus

Barrett's esophagus (BE) is a medical condition where the normal mucosal cells lining in the lower portion of the esophagus is replaced with acid-resistant cells that are normally only present in the small and large intestine. BE is considered a premalignant condition since the disease is associated with a high incidence of the often-deadly cancer, esophageal adenocarcinoma (EAC). Patients suffering from BE have a 30-fold increased chance of developing EAC. When detected at an early stage, cancer in BE can still be endoscopically treated and has an excellent

prognosis [23]. However, in many cases, early stages of EAC go undetected as routine endoscopic procedures are performed by endoscopists who are unfamiliar with the early visual signs of the disease [24]. Therefore, the recommended protocol entails taking evenly spaced biopsies throughout the esophagus. Unfortunately, studies have shown that the current biopsy protocol is prone to sampling error which leads to missed lesions. As a result, relatively easy-to-treat early cancer develops into late-stage cancer where invasive surgery and radiochemotherapy are the only available treatment options. In the following sections, we will discuss the use of AI systems in Barrett's esophagus. A succinct overview of relevant references is shown in Table 1.

BE Cancer Detection Using White Light Endoscopy

Prior to the deep learning revolution, many researchers attempted to classify dysplasia in Barrett's esophagus using conventional machine learning techniques. Van der Sommen et al. [25] used Gabor filters to extract features from patches of BE images in overview. These features were then used to train a support vector machine (SVM) classifier to make predictions on new patches. A segmentation could then be created from combining the patch scores. Similarly, Souza et al. [26] also employed an SVM to classify dysplasia in BE, but extracted features using color co-occurrence matrices. Others attempt to bridge the gap between deep learning and a handcrafted feature approach by incorporating both in the same model. In the work by Ghatwary et al. [27], Gabor features were combined with features extracted from a DenseNet architecture [28] to perform classification on the Kvasir dataset [29] and a dataset from an endoscopic vision challenge [30]. Van Riel et al. [31] did employ CNNs but used so-called CNN codes. With this approach, a pretrained CNN (generally pretrained using the ImageNet dataset) is used, where features are extracted from the penultimate layer. In this work, Alexnet [32], VGG [33], and GoogleNet [34] were used as feature extractors, and SVM and

Table 1 Overview of the relevant references – C: Classification, S: Segmentation, D: Detection, R: Reconstruction, CNN: Convolutional neural networks, SSD: Single shot detector, EOCT: Endoscopic optical coherence tomography, CAM: Class activation map, RNN: Recurrent Neural Network, CV: Cross validation, LOO: Leave one out,

References					
WLE	Task	Description	Validation type	Modality	Dataset
Van der Sommen 2016 [25]	C + S	Gabor features and SVM	LOO-CV	WLE	EndoVis
Souza 2018 [26]	C	Color co-occurrence and SVM	LOPO-CV	WLE + NBI	EndoVis + Ausburg (in house)
Riel 2018 [31]	C	CNN codes	LOPO-CV	WLE	EndoVis
Ghatwary 2019 [27]	C	CNN + Gabor	LOPO-CV	WLE	EndoVis + Kvasar
Mendel 2017 [36]	C	Fine-tune Resnet	LOPO-CV	WLE + NBI	EndoVis
Ebigbo 2020 [38]	C + S	Resnet + DeeplabV3+	Clinical	WLE	–
Souza 2020 [42]	C	GAN data augmentation	Tr/Val	WLE	EndoVis + Ausburg (in house)
Ghatwary 2019-II [43]	C + D	R-CNN and SSD	LOPO-CV	WLE	EndoVis
Ghatwary 2020 [48]	C + D	R-CNN and SSD	Tr/Val/Te	WLE	EndoVis
Hashimoto 2020 [49]	C + D	Xception + Yolo	Tr/Val	WLE + NBI	In-house dataset
Van der Putten 2019 [51]	C + S	Domain specific pretraining and CNN	Tr/Val/Te	WLE	In-house dataset
De Groof 2020 [50]	C + S	U-Net/Resnet hybrid	Tr/Val/Te	WLE	In-house dataset
Van der Putten 2020 [52]	C + S	U-Net/Resnet hybrid	Clinical	WLE	–
De Groof 2020-II [53]	C + S	Clinical evaluation of developed model	Clinical	WLE	–
Van der Putten 2020-II [54]	C	Endoscopic video classification	PP-CV	WLE	In-house dataset
Verhage 2020 [55]	C	Field effect classification	PP-CV	WLE	In-house dataset
NBI	Task	Description	Validation type	Modality	Dataset
Rajan 2009 [63]	C	Multimodality classification with SVM/KNN	CV WLE + NBI + AAC		In-house dataset
Van der Putten 2019-II [64]	C	NBI-zoom video characterization using Resnet model	PP-CV	WLE + NBI	In-house dataset
Struyvenberg 2020 [66]	C	NBI-zoom video characterization using Resnet model	PP-CV	NBI	In-house dataset
Endomicroscopy	Task	Description	Validation type	Modality	Dataset
Grisan 2012 [68]	C	Two-stage classification with feature extraction	LOO-CV	pCLE	In-house dataset
Veronese 2013 [69]	C	Hybrid image and patch-based feature extraction	LOO-CV	pCLE	In-house dataset

(continued)

Table 1 (continued)

References					
Hong 2017 [70]	C	CNN for classification of endomicroscopy images	Tr/Val	pCLE	ISBI 2016 challenge
Ghatwary 2017 [71]	C	Feature extraction + SVM for endomicroscopy images	LOO-CV	pCLE	In-house dataset
Ghatwary 2019-III [72]	C	Feature enhancement + SVM	LOO-CV	pCLE	In-house dataset
Pulido 2020 [73]	C	BE screening using endomicroscopy videos	Tr/Val	pCLE	In-house dataset
Qi 2004 [75]	C	EOCT classification using CSAC	LOO-CV	EOCT	In-house dataset
Qi 2006 [76]	C	EOCT classification using Fourier analysis	LOO-CV	EOCT	In-house dataset
Rodriguez 2015 [77]	C	GLCM + bayes classifier	LOO-CV	VLE	In-house dataset
Swager 2017 [96]	C	Clinically inspired classification of VLE	CV	VLE	In-house dataset
Scheeve 2019 [78]	C	Clinically inspired glands feature classification	Tr/Val/Te	VLE	In-house dataset
Van der Sommen 2018 [79]	C	Multiple handcrafted features + extensive classifier analysis	LOO-CV	VLE	In-house dataset
Putten 2019-III [80]	S	Region of interest segmentation with U-Net	Fivefold CV	VLE	In-house dataset
Fonalla 2019 [81]	C	Ensemble of CNNs + CAM	Tr/Val/Te	VLE	In-house dataset
Putten 2020-III [82]	C	Principal dimension encoding classification	Fivefold CV	VLE	In-house dataset
Quality assessment	Task	Description	Validation type	Modality	Dataset
Hwang 2005 [84]	C	GLCM + SURF bubble classification	Twofold CV	WLE	In-house dataset
Pietri 2018 [86]	C	DFT + frame clustering	–	WLE	In-house dataset
Wang 2019 [85]	C	RSS filter	Tr	WLE	In-house dataset
Akbari 2018 [87]	C + S	HSV + RGB features	Tr	WLE	CVC-ColonDB
Van Dongen 2016 [88]	C	DCT + color + blur-based features	Tenfold CV	WLE	In-house dataset
Tajbakhsh 2016 [89]	C	Full training versus fine-tuning	Tr/Val	WLE	In-house dataset
Islam 2018 [90]	C	Bubble, water, and blurry frame detection + CNN	Tr/Val	WLE	In-house dataset
Hong 2014 [91]	R	3D colon reconstruction with depth from intensity	–	WLE	In-house dataset

(continued)

Table 1 (continued)

References					
Putten 2019-IV [92]	C	Informative frame classification with CNN + HMM	Fivefold CV	WLE	In-house dataset
Boers 2020 [93]	C	Three informative frame classification with RNN	Fivefold CV	WLE	In-house dataset
Boers 2020-II [94]	C	Tissue classification with RNN	Fivefold CV	WLE	In-house dataset

random forest [35] were used to further classify the images.

Mendel et al. published the first preliminary results for adenocarcinoma detection using CNNs in the esophagus [36]. In this work, the same endoscopic vision challenge dataset was used by fine-tuning a ResNet model [37] on patches of the full-resolution images. The same group has done more research on WLE images using deep learning since then. In their most recent deep learning WLE study, Ebigo et al. [38] developed a new algorithm with more advanced deep learning techniques, such as spatial pyramid pooling [39], and a state-of-the-art decoder [40]. It is noteworthy that the algorithm is not trained and tested on WLE images in overview but on zoomed in lesions/healthy tissue. Therefore, this algorithm cannot be used for primary detection, however the system shows excellent performance with an overall accuracy of 89.9%. In more recent work of this group, the use of Generative Adversarial Networks (GAN) [41] is proposed to increase the variation of data augmentation [42]. Fake images of Barrett's esophagus are generated alongside real examples in order to increase the variation in the dataset. Results showed that generating artificial patches of BE increases the classification performance compared to using the original dataset alone.

A different group performed an extensive evaluation of many existing deep learning architectures for the detection of early esophageal adenocarcinoma. In work by Ghatwary et al. [43], several object detection frameworks such as Regional-Based Convolutional Neural Networks [44] (R-CNN), Fast R-CNN [45], Faster R-CNN [46], and Single-Shot Multibox Detector

(SSD) [47] are implemented and evaluated on the data from the endoscopic vision challenge. The results from this work show that even with limited data, deep learning can outperform traditional machine learning methods for the detection of cancer. The same group also recently proposed a system for the automatic detection of endoscopic abnormalities from endoscopic video [48]. In this work, a 3D CNN is combined with a convolutional long short-term memory (LSTM) to efficiently learn short- and long-term spatio-temporal features. The learned feature map is then utilized by a Region Proposal Network in order to produce bounding boxes that indicate cancer in the videos. The aforementioned publications regarding endoscopic cancer detection in BE employ relatively little data compared to other deep learning research, which generally perform better with more data. In 2020, Hashimoto et al. [49] published a study with considerably more data (nearly 1,000 images) compared to earlier studies by other groups that only had access to smaller datasets (several hundred). The dataset contained both WLE and NBI images (including zoom imagery). A CNN algorithm was pretrained (on ImageNet) and subsequently fine-tuned to perform binary classification. An additional object detects on algorithm with the goal of drawing bounding boxes surrounding the dysplastic regions. An impressive result of 95% accuracy for classification and mean-average precision of 0.7533 for detection was obtained with real-time performance. However, the algorithm was trained on a limited amount of patients (100 total).

De Groof et al. [50] and Van der Putten et al. [51] recently demonstrated the benefits of using a

very large dataset of endoscopic images ($\approx 500,000$) to *pretrain* a convolutional neural network. Subsequently, a several BE-specific datasets are used to fine-tune the model, each time getting closer to the target domain. They compared their CAD results to 53 physicians in four different skill levels and the results show that the AI algorithm outperforms every doctor by a convincing margin [50]. In later studies, the same algorithm is evaluated live in the clinic with excellent results where 90% of patients get correctly classified by the algorithm [52, 53]. The same group also investigated challenges and opportunities of AI with respect to endoscopic video [54]. A CNN trained on WLE images of EAC was tasked to find lesions in endoscopic video as well. While the results are promising, noninformative frames prove to be a problem. A second problem is how to determine the diagnosis of an entire video when only a small section is actually cancerous, false positives make this a difficult problem. Lastly, a study was conducted into the so-called “field effect,” which theorizes that nondysplastic tissue in dysplastic BE exhibits similar characteristics that are not as obvious as the actual lesion, but might be perceptible by AI. In a study by Verhage et al. [55], a CNN is employed on nondysplastic patches of dysplastic images and compared to patches of a nondysplastic Barrett's esophagus. Results show a statistical difference between the two which indicates the potential presence of a visual field effect.

BE Cancer Detection Using Narrow-Band Imaging

While white light endoscopy is the de facto choice for endoscopic imaging [56], technological advances have led to several alternative imaging modalities with the aim of improving cancer detection and lesion characterization. Narrow-band imaging (NBI) is based on narrowing the spectral transmittance of white light using optical filters with the aim of improving the sensitivity of detecting superficial esophageal cancer [57]. This is often combined with the zoom modality of newer endoscopes aiming to characterize the

cancer (as opposed to primary detection for WLE). Several NBI (non-AI) classification schemes have been proposed in the last decade using a variety of criteria but are still suboptimal [58–60]. Recently, the combination of artificial intelligence and narrow band imaging has been used extensively in other closely related fields such as colonoscopy [10, 61, 62]. For BE, however, the use of NBI data is limited.

In 2009, Rajan et al. [63] used WLE, NBI, and a third modality called acetic acid chromoendoscopy (AAC). Washes of acetic acid are commonly used in colonoscopy to highlight dysplastic areas and thereby enhance the ability to obtain targeted biopsy specimens. This technique can also be used to identify small islands of residual specialized columnar epithelium and thus enhance mucosal pit patterns during endoscopic imaging. In their paper, three classification algorithms (SVM, K-nearest neighbors, and boosting) are used in combination with the three mentioned modalities to classify lesions. Results from the study show that NBI and AAC outperform WLE on their dataset.

NBI can also be combined with WLE to maximize the information extracted from both modalities. Van der Putten et al. [64] uses a small paired dataset of WLE and NBI images (images were taken sequentially with as little camera movement as possible). Regular affine registration was employed to ensure a better match between the two modalities. Two CNNs were trained on the two modalities and results showed that a composite prediction increased the localization performance of the algorithm when compared to only the WLE algorithm. Later, the same group performed a characterization study on a NBI video dataset [65, 66]. The CNN was pretrained on a large endoscopy-specific dataset (same dataset as in [50]) and trained on 183 NBI-zoom videos. The results showed that endoscopic pre-training outperformed natural image pretraining (ImageNet) as well as training from scratch. Additionally, the algorithm was able to run in real time.

Other studies mentioned previously ([26, 36, 49]) also employed NBI images in their datasets. However, in these studies, NBI data was either added to the training data or evaluated separately.

NBI data was not used as a means to enhance the WLE predictions in these cases.

BE Cancer Detection Using Endomicroscopy

Due to the difficulty in detecting early cancer using conventional optical methods such as WLE and NBI, new methods have been developed based on long-wavelength lasers that can penetrate the skin and visualize the subsurface tissue in order to potentially find early lesions in these tissue layers. Endomicroscopy is a technique for obtaining histology-like images from inside the human body in real time, a process known as “optical biopsy.” These methods are unique in the sense that they permit *in vivo* histologic analysis of esophageal mucosa, which is not possible with traditional methods.

1. Probe-based confocal laser endomicroscopy:

Probe-based confocal laser endomicroscopy (pCLE) comprises a fiberoptic bundle that can be inserted into the accessory channel of modern endoscopes. The probe-based system has a fixed focal length, limiting the scan to a single plane. Probe-based CLE is able to obtain very high-magnification and high-resolution images of the mucosal layer of the GI tract. The system operates by detecting reflections of the fluorescence of light through a pinhole [67]. While research has shown that pCLE can lead to an increased specificity for identifying high-grade dysplasia, the associated sensitivity is still low compared to conventional methods.

Grisan et al. [68] introduced a computer diagnosis method in 2012 for classification between intestinal metaplasia (IM), gastric metaplasia (GM), and neoplasia mucosa (NPL) with an overall accuracy of 84%. Feature extraction was performed on an image basis and further processed with a two-stage classifier. First, images were classified as NPL or not. Second, the non-NPL images were further classified as either IM or GM. Leave-one-out cross-validation was used to evaluate performance on 336 CLE images. Later,

Veronese et al. [69] proposed a hybrid approach of image and patch-based feature extraction using intensity distribution values, geometric characteristics, and local binary patterns to determine if an image is IM or not in a similar two-stage approach as prior work [68].

In 2017, Hong et al. [70] investigated whether machine learning models may improve screening accuracy via pCLE and recognize features that may have eluded human visual analysis. A small CNN consisting of four convolutional layers and two max-pooling layers was followed by several fully connected layers. While results look promising, the imbalance in size of different subcategories caused low accuracy for predicting the minority class.

Ghatwary et al. [71] proposed a classification model that discriminated between the IM, GM, and NPL on CLE images. The model was divided into three phases: the first phase was improving the image details by using an ad hoc filter and the second phase was extracting different features, such as intensity features, wavelet features, gray-level co-occurrence matrix (GLCM), fractal dimension, and fuzzy LBP. Later, aiming to improve this result, Ghatwary et al. [72] proposed to use a specific preprocessing filter followed by traditional machine learning techniques. The extracted features were used to train a model, which was subsequently tested on a dataset consisting of 557 images of four different histopathology grades from 96 patients. The model obtained excellent scores (accuracy of 96%) and showed that preprocessing with a discrete wavelet transform improved model performance.

In work by Pulido et al. [73], pCLE *videos* are classified into the same three classes as in [70]. Two video models are introduced that classify frames from a dataset compiled from 78 patients with 10–15 videos per patient examining different areas of the esophagus. Results show high specificity and increased sensitivity when using the proposed model.

2. *Optical coherence tomography:* Optical coherence tomography (OCT) is an emerging optical technique using near-infrared light to generate

high-resolution images of tissue structures, providing noninvasive, subsurface, high-resolution imaging of biological microstructure [74]. The conceptual idea of OCT is comparable to ultrasound, but instead of using the reflection of acoustic waves, OCT uses the scattering of near-infrared light to generate images. Subsequently, consecutive 2D images generated by pullbacks can be reconstructed to 3D structures.

The work by Qi et al. [75] has shown promising results for the detection of dysplasia using endoscopic OCT and texture analysis. Center-symmetric autocorrelation features such as gray-scale texture covariance, local variance, between-pair variance, and within-pair variance were calculated and PCA was used to obtain classification results. An AUC of 84% was obtained with this method. Later, the same group improved on their previous result by using more complex Fourier domain fractal analysis and classification trees [76]. Perfect classification scores (100% accuracy) were achieved with the proposed method.

3. *Volumetric laser endomicroscopy*: Volumetric laser endomicroscopy (VLE) is a second-generation OCT application with a significantly higher image quality compared to standard OCT. The tissue penetration depth of the VLE scan is 3 mm with an axial resolution of $7\text{ }\mu\text{m}$ and a lateral resolution of $40\text{ }\mu\text{m}$. Even though endoscopic OCT can ensure histopathology correlation between the scan and the tissue, only a small portion can be imaged at a time. In contrast, VLE can capture the complete circumferential and longitudinal Barrett segment in a single scan with a much higher image quality. In 2015, Rodriguez-Diaz and Singh et al. [77] presented an algorithm for image interpretation of VLE images using gray-level co-occurrence matrix statistics of the first wavelet components of the image, followed by a naive Bayes classifier. An impressive sensitivity and specificity of 86% and 93%, respectively, were reported for computer-aided classification between dysplastic and nondysplastic BE tissue. Swager et al. presented a computer-aided detection

algorithm where two clinically inspired quantitative image features specific to VLE were developed based on the VLE surface signal and the intensity histogram of several layers. Scheeve et al. [78] also employed a clinically inspired approach where quantitative image features were developed based on gland patterns in the submucosal tissue visualized by VLE. In 2018, Van der Sommen et al. [79] investigated a large variety of imaging features such as GLCM, local binary patterns, and histogram features. Multiple classifiers such as SVM, random forest, KNN, and logistic regression were used to assess the discriminative power of the extracted features. While results were promising, the tissue of interest had to be extracted by hand prior to classification. Work by Van der Putten et al. [80] attempts to solve this problem by creating a deep learning-based segmentation model. The decreasing signal-to-noise ratio in deeper layers of tissue was used to improve the segmentation performance. Resulting segmentations were indistinguishable from experts when compared to interobserver variability of different experts. Fonolla et al. [66, 81] employed an ensemble of CNN models to obtain an AUC of 0.96 on an independent test set in a multicenter study. The study also investigates the most relevant regions of the image by using class activation maps. Finally, Van der Putten et al. [82] proposed principal dimension encoding as a means to incorporate the inherent scanning properties of the VLE modality to improve classification results. First, the region of interest segmentations from earlier work [80] was used to extract relevant tissue. Then, scan lines were encoded one by one prior to classifying with several classification algorithms such as SVM and random forest.

AI for Quality Assessment in the Esophagus

The vast majority of AI research in the gastrointestinal (GI) tract is primarily focused on detection of diseases, often using heavily curated datasets.

While this type of research can lead to algorithms that are extremely useful in clinical practice, the problem remains that you cannot classify what is not properly visualized. In recent years, several algorithms have attempted to improve the quality of endoscopies by detecting informativeness of frames, detecting specific artifacts, or assessing missed regions in the esophagus. For clarity, not all references in this section pertain to the esophagus exclusively, quality assessment of videos and images in other regions of the gastrointestinal tract such as the small intestine are applicable to BE as well in this case.

Early work for the quality assessment in the GI tract was based on traditional feature engineering. In work by Hwang et al. [83], blurry frames are classified based on the number of discontinuing edges on that frame, motivated by the observation that a sharp image typically has fewer isolated edges compared to a blurry image. In later work [84], the same group improved on the classifier by computing eigenvalues of the Fourier spectrum of the frames and using K-means to detect blurry frames. A different group [85] uses a handcrafted “Ring Shape Selective” filter to detect ring shape bubbles in the image. Similarly, Pietri et al. [86] detect bubbles in still frames with four traditional features. For specular reflection detection, Akbari et al. [87] combined RGB and HSV color space features to detect highlights in an image with an SVM classifier. For BE specifically, Van Dongen et al. [88] distinguish frames between informative and noninformative based on a variety of features. Discrete cosine transform-based features, color features, and blur-based features are used in combination with a regression tree to classify frames.

As with most other medical imaging applications, deep learning has also been extensively used for artifact detection in the GI tract. In 2016, Tajbakhsh et al. [89] performed a large study to determine whether fine-tuning outperforms full training. In one of the experiments, informativeness of endoscopic video frames was analyzed and showed that the pretrained network using a large set of labeled natural images provides a better accuracy than a CNN trained from scratch. Islam et al. [90] used CNNs to classify noninformative frames in colonoscopy videos as well. In this study, distinctions were made

between blurry images, and images containing water or bubbles. A very efficient architecture was proposed that traded processing speed for a small decrease in performance. Other methods focus on reconstructing a virtual colon from colonoscopy frames to ensure all areas of the colon have been visualized [91]. A depth-from-intensity technique was proposed to estimate the distance of each fold from the camera using the brightness intensity of selected pixels around fold contours.

For Barrett’s esophagus specifically, Van der Putten et al. [92] employed a CNN to classify individual video frames of BE pullbacks as either informative or noninformative. A hidden Markov model was subsequently used to incorporate temporal information, which helped to improve results. Later, the same group improved on their work by replacing the hidden Markov model by a recurrent neural network [93]. This allows training with video sequences end to end instead of two stages. In other related work, Boers et al. [94] employed a similar recurrent model to detect time sequences in endoscopic pullbacks which contained tissue that could potentially contain Barrett’s cancer with the aim to decrease the amount of false positives.

The body of work on quality assessment in endoscopic video is very large and we only cover a subset of all available publications. For a more complete review of all existing work in this field, we refer to a recent review of De Groen [21], which addresses different aspects that need to be considered when using artificial intelligence to improve adequacy of inspection in gastrointestinal endoscopy which goes deeper into this aspect.

Discussion

The current sampling protocol for surveillance Barrett’s endoscopies is error prone and heavily dependent on the skill of the endoscopist. For this reason, artificial intelligence that can mitigate these problems could be extremely valuable in the fight against esophageal cancer. The technologies that have emerged so far have attempted to replace random biopsies that often claim extremely high accuracies on in-house datasets or highly curated datasets with a nonnegligible

selection bias. However, performance in clinical practice is often never tested. While pilot studies and initial experiments are a natural part of the development process, clinical studies with new images/videos from unseen patients are necessary to properly validate how these algorithms translate to the clinic. Fortunately, there seems to be a paradigm shift in the literature placing higher emphasis on proper validation [22, 95]. The most recent studies [38, 52, 53] have also started testing promising algorithms in the clinic. This is a vital step where interdisciplinary knowledge and resources are of paramount importance.

Other modalities such as NBI and VLE have shown to be beneficial or complimentary to existing WLE algorithms. While the reported performance of these methods is extremely good, none have been evaluated in live endoscopies so far. This might be explained by the fact that white light endoscopy is the standard imaging modality and available to all endoscopists. NBI and VLE on the other hand are much rarer and require expensive specialized equipment. Considering that, when aiming for an improved detection, these methods are to be mainly used in periphery hospitals, sufficient funds might not be available to purchase these devices. This limits the applicability of these techniques, especially if standard white light endoscopy provides similar performance.

Given that dysplasias in Barrett's esophagus are nonpolypoid, a higher skill level is required of the endoscopists in order to achieve a good detection rate. Unfortunately, most practicing endoscopists have suboptimal detection and miss lesions despite being detectable by an expert endoscopist. A second problem is the relatively low prevalence of BE dysplasia, which inhibits an extensive training program. If the performance (90 + % accuracy) of the models in the clinical pilot studies prove to hold true, most endoscopists in peripheral centers will likely perform much better when assisted by AI. Another factor that compounds the problem is that endoscopists tend to specialize in one specific disease according to the prevalence specific to the country. In contrast, given a dataset of sufficient size, AI can be trained to classify any number of diseases. This is showcased by high accuracies on a large amount of different diseases that use very similar models (CNN with residual connections).

The research on quality assessment of endoscopies is broadly acknowledged to be very promising. Many of the proposed algorithms present high accuracies for either detecting non-informative frames in a video sequence or detection of unwanted artifacts such as bubbles or specular reflections in endoscopic images. Interestingly, no studies have been performed so far that combine the quality algorithms with cancer detection algorithms to improve cancer detection performance. Additionally, to the best of our knowledge, these algorithms have also not been tested in the clinic to improve cancer detection by endoscopists. One opportunity for future work would be to combine quality assessment algorithms with cancer detection algorithms that can be trained and evaluated on data that is not heavily curated. A second opportunity for future work that would have tremendous value would be to use quality assurance algorithms in the clinic with nonexpert endoscopists to assess whether their detection rate improves compared to nonexperts that do not use these algorithms. A third avenue that has not yet been explored is the effect that AI has on the user. On the one hand, human performance may improve by the mere fact of being watched by AI (the Hawthorne effect), on the other hand the sense that AI is watching and there is less need to carefully inspect the tissue may cause the endoscopist to produce suboptimal videos. Unfortunately, even the best AI cannot compensate for poor input quality.

In this review, developments of AI in Barrett's esophagus have been discussed. While it is evident that a lot of research is focused on computer-aided detection and segmentation of cancer in BE, most algorithms are not ready for clinical implementation. Regardless, the current research is very promising and developments in the field will provide many opportunities for the future of AI in Barrett's esophagus and ideally improve patient care.

References

1. Arnold M, Soerjomataram I, Ferlay J, Forman D. Global incidence of oesophageal cancer by histological subtype in 2012. *Gut*. 2015;64(3):381–7.
2. Schölvinck DW, Van Der Meulen K, Bergman JJ, Weusten BL. Detection of lesions in dysplastic

- Barrett's esophagus by community and expert endoscopists. *Endoscopy*. 2017;49(02):113–20.
3. Enzinger PC, Mayer RJ. Esophageal cancer. *N Engl J Med*. 2003;349(23):2241–52.
 4. Cotton CC, Wolf WA, Overholt BF, Li N, Lightdale CJ, Wolfsen HC, Pasricha S, Wang KK, Shaheen NJ, Sampliner RE, et al. Late recurrence of Barrett's esophagus after complete eradication of intestinal metaplasia is rare: final report from ablation in intestinal metaplasia containing dysplasia trial. *Gastroenterology*. 2017;153(3):681–8.
 5. Phoa KN, Rosmolen WD, Weusten BL, Bisschops R, Schoon EJ, Das S, Ragunath K, Fullarton G, DiPietro M, Ravi N, et al. The cost-effectiveness of radiofrequency ablation for barrett's esophagus with low-grade dysplasia: results from a randomized controlled trial (surf trial). *Gastrointest Endosc*. 2017;86(1):120–9.
 6. Naveed M, Kubilun N. Endoscopic treatment of early-stage esophageal cancer. *Curr Oncol Rep*. 2018;20(9):71.
 7. Syed T, Doshi A, Guleria S, Syed S, Shah T. Artificial intelligence and its role in identifying esophageal neoplasia. *Dig Dis Sci*. 2020;1–8.
 8. Bejnordi BE, Veta M, Van Diest PJ, Van Ginneken B, Karssemeijer N, Litjens G, Van Der Laak JA, Hermans M, Manson QF, Balkenhol M, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*. 2017;318(22):2199–210.
 9. Ghafoorian M, Karssemeijer N, Heskes T, van Uden IW, Sanchez CI, Litjens G, de Leeuw F-E, van Ginneken B, Marchiori E, Platel B. Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. *Sci Rep*. 2017;7(1):1–12.
 10. Byrne MF, Chapados N, Soudan F, Oertel C, Pérez ML, Kelly R, Iqbal N, Chandelier F, Rex DK. Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. *Gut*. 2019;68(1):94–100.
 11. Mori Y, Kudo S-e, Misawa M, Saito Y, Ikematsu H, Hotta K, Ohtsuka K, Urushibara F, Kataoka S, Ogawa Y, et al. Real-time use of artificial intelligence in identification of diminutive polyps during colonoscopy: a prospective study. *Ann Intern Med*. 2018;169(6):357–66.
 12. van der Sommen F, Curvers WL, Nagengast WB. Novel developments in endoscopic mucosal imaging. *Gastroenterology*. 2018;154(7):1876–86.
 13. Pech O, Bollschweiler E, Manner H, Leers J, Ell C, Hölscher AH. Comparison between endoscopic and surgical resection of mucosal esophageal adenocarcinoma in Barrett's esophagus at two high-volume centers. *Ann Surg*. 2011;254(1):67–72.
 14. Gavin A, Francisci S, Foschi R, Donnelly D, Lemmens V, Brenner H, Anderson L, E.- W. Group, et al. Oesophageal cancer survival in europe: a eurocare-4 study. *Cancer Epidemiol*. 2012;36(6):505–12.
 15. Njei B, McCarty TR, Birk JW. Trends in esophageal cancer survival in United States adults from 1973 to 2009: a seer database analysis. *J Gastroenterol Hepatol*. 2016;31(6):1141–6.
 16. Das A, Singh V, Fleischer DE, Sharma VK. A comparison of endoscopic treatment and surgery in early esophageal cancer: an analysis of surveillance epidemiology and end results data. *Am J Gastroenterol*. 2008;103(6):1340–5.
 17. Rice TW, Ishwaran H, Ferguson MK, Blackstone EH, Goldstraw P. Cancer of the esophagus and esophagogastric junction: an eighth edition staging primer. *J Thorac Oncol*. 2017;12(1):36–42.
 18. Sharma N, Ho KY. Recent updates in the endoscopic diagnosis of Barrett's oesophagus. *Gastrointest Tumors*. 2016;3(2):109–13.
 19. Shaheen NJ, Falk GW, Iyer PG, Gerson LB. ACG clinical guideline: diagnosis and management of Barrett's esophagus. *Am J Gastroenterol*. 2016;111(1):30–50.
 20. Wani S, Williams JL, Komanduri S, Muthusamy VR, Shaheen NJ. Endoscopists systematically undersample patients with longsegment Barrett's esophagus: an analysis of biopsy sampling practices from a quality improvement registry. *Gastrointest Endosc*. 2019;90(5):732–41.
 21. de Groen PC. Using artificial intelligence to improve adequacy of inspection in gastrointestinal endoscopy. *Tech Gastrointest Endosc*. 2019;22(2):150640.
 22. F. van der Sommen, J. de Groof, M. Struyvenberg, J. van der Putten, T. Boers, K. Fockens, E. J. Schoon, W. Curvers, Y. Mori, M. Byrne, et al. Machine learning in GI endoscopy: practical guidance in how to interpret a novel field. *Gut*, 2020;69:2035–2045.
 23. Pech O, May A, Manner H, Behrens A, Pohl J, Weferling M, Hartmann U, Manner N, Huijsmans J, Gossner L, et al. Long-term efficacy and safety of endoscopic resection for patients with mucosal adenocarcinoma of the esophagus. *Gastroenterology*. 2014;146(3):652–60.
 24. Boerwinkel DF, Swager A-F, Curvers WL, Bergman JJ. The clinical consequences of advanced imaging techniques in Barrett's esophagus. *Gastroenterology*. 2014;146(3):622–9.
 25. van der Sommen F, Zinger S, Curvers WL, Bisschops R, Pech O, Weusten BL, Bergman JJ, Schoon EJ, et al. Computer-aided detection of early neoplastic lesions in Barrett's esophagus. *Endoscopy*. 2016;48(07):617–24.
 26. Souza L, Ebigo A, Probst A, Messmann H, Papa JP, Mendel R, Palm C. Barrett's esophagus identification using color co-occurrence matrices. In: 2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI). Los Alamitos: IEEE; 2018. p. 166–73.
 27. Ghatwary N, Ye X, Zolgharni M. Esophageal abnormality detection using densenet based faster r-cnn with gabor features. *IEEE Access*. 2019;7:84374–85.

28. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 4700–8.
29. Pogorelov K, Randel KR, Griwodz C, Eskeland SL, de Lange T, Johansen D, Spampinato C, Dang-Nguyen D-T, Lux M, Schmidt PT, et al. Kvasir: a multi-class image dataset for computer aided gastrointestinal disease detection. In: Proceedings of the 8th ACM on multimedia systems conference; Association for Computing Machinery, New York; 2017. p. 164–9.
30. Sub-challenge early barrett's cancer detection. NA. <https://endovissub-barrett.grand-challenge.org>. 2017.
31. Van Riel S, Van Der Sommen F, Zinger S, Schoon EJ, de With PH. Automatic detection of early esophageal cancer with cnns using transfer learning. In: 2018 25th IEEE international conference on image processing (ICIP). IEEE; 2018. p. 1383–7.
32. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84–90.
33. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014.
34. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition; IEEE, Boston; 2015. p. 1–9.
35. Ho TK. Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition, vol. 1. Los Alamitos: IEEE; 1995. p. 278–82.
36. Mendel R, Ebigbo A, Probst A, Messmann H, Palm C. Barrett's esophagus analysis using convolutional neural networks. In: Bildverarbeitung für die Medizin. Berlin: Springer; 2017. p. 80–5.
37. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Los Alamitos: IEEE; 2016. p. 770–8.
38. Ebigbo A, Mendel R, Probst A, Manzeneder J, Prinz F, de Souza Jr LA, Papa J, Palm C, Messmann H. Real-time use of artificial intelligence in the evaluation of cancer in barrett's oesophagus. Gut. 2020;69(4):615–6.
39. He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell. 2015;37(9):1904–16.
40. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H. Encoderdecoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV), Springer, Munich; 2018. p. 801–18.
41. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Advances in neural information processing systems, Curran Associates Inc, Montreal; 2014. p. 2672–80.
42. de Souza Jr LA, Passos LA, Mendel R, Ebigbo A, Probst A, Messmann H, Palm C, Papa JP. Assisting barrett's esophagus identification using endoscopic data augmentation based on generative adversarial networks. Comput Biol Med. 2020;126:104029.
43. Ghatwary N, Zolgharni M, Ye X. Early esophageal adenocarcinoma detection using deep learning methods. Int J Comput Assist Radiol Surg. 2019;14(4):611–21.
44. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway: IEEE; 2014. p. 580–7.
45. Girshick R. Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. Piscataway: IEEE; 2015. p. 1440–8.
46. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, Curran Associates Inc, Montreal; 2015. p. 91–9.
47. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC. Ssd: Single shot multibox detector. In: European conference on computer vision. Berlin: Springer; 2016. p. 21–37.
48. Ghatwary N, Zolgharni M, Janan F, Ye X. Learning spatiotemporal features for esophageal abnormality detection from endoscopic videos, vol. 25. IEEE J Biomed Health Inform; 2020. p. 131–42.
49. Hashimoto R, Requa J, Tyler D, Ninh A, Tran E, Mai D, Lugo M, Chehade NE-H, Chang KJ, Karnes WE, et al. Artificial intelligence using convolutional neural networks for real-time detection of early esophageal neoplasia in barrett's esophagus (with video), vol. 91. Gastrointest Endosc; 2020. p. 1264–1271.e1.
50. de Groof AJ, Struyvenberg MR, van der Putten J, van der Sommen F, Fockens KN, Curvers WL, Zinger S, Pouw RE, Coron E, Baldaque-Silva F, et al. Deep-learning system detects neoplasia in patients with barrett's esophagus with higher accuracy than endoscopists in a multistep training and validation study with benchmarking. Gastroenterology. 2020;158(4):915–29.
51. van der Putten J, de Groof J, van der Sommen F, Struyvenberg M, Zinger S, Curvers W, Schoon E, Bergman J, et al. Pseudo-labeled bootstrapping and multi-stage transfer learning for the classification and localization of dysplasia in barrett's esophagus. In: International workshop on machine learning in medical imaging. Cham: Springer; 2019. p. 169–77.
52. van der Putten J, de Groof J, Struyvenberg M, Boers T, Fockens K, Curvers W, Schoon E, Bergman J, van der Sommen F, de With PH. Multi-stage domain-specific pretraining for improved detection and localization of barrett's neoplasia: a comprehensive clinically validated study. Artif Intell Med. 2020;107:101914.
53. de Groof AJ, Struyvenberg MR, Fockens KN, van der Putten J, van der Sommen F, Boers TG, Zinger S,

- Bisschops R, Peter H, Pouw RE, et al. Deep learning algorithm detection of barrett's neoplasia with high accuracy during live endoscopic procedures: a pilot study (with video). *Gastrointest Endosc.* 2020;91: 1242–50.
54. van der Putten J, de Groot J, van der Sommen F, Struyvenberg M, Zinger S, Curvers W, Schoon E, Bergman J, et al. First steps into endoscopic video analysis for barrett's cancer detection: challenges and opportunities. In: Medical imaging 2020: computer-aided diagnosis, vol. 11314. International Society for Optics and Photonics, Washington; 2020. p. 1131431.
55. Verhage L, van der Putten J, van der Sommen F, de Groot J, Struyvenberg M, et al. The field effect in barrett's esophagus: a macroscopic view using white light endoscopy and deep learning. In: Medical imaging 2020: computer-aided diagnosis, vol. 11314. International Society for Optics and Photonics, Washington; 2020. p. 1131437.
56. Wei W-Q, Chen Z-F, He Y-T, Feng H, Hou J, Lin D-M, Li X-Q, Guo C-L, Li S-S, Wang G-Q, et al. Long-term follow-up of a community assignment, one-time endoscopic screening study of esophageal cancer in china. *J Clin Oncol.* 2015;33(17):1951.
57. Sano Y. New diagnostic method based on color imaging using narrowband imaging (NBI) system for gastrointestinal tract. *Gastrointest Endosc.* 2001;53: AB125.
58. Kara MA, Ennahachi M, Fockens P, ten Kate FJ, Bergman JJ. Detection and classification of the mucosal and vascular patterns (mucosal morphology) in Barrett's esophagus by using narrow band imaging. *Gastrointest Endosc.* 2006;64(2):155–66.
59. Nogales O, Caballero-Marcos A, Clemente-Sánchez A, García-Lledó J, Pérez-Carazo L, Merino B, Carbonell C, López-Ibáñez M, González-Asanza C. Usefulness of non-magnifying narrow band imaging in evis exera iii video systems and high-definition endoscopes to diagnose dysplasia in barrett's esophagus using the Barrett international NBI group (bing) classification. *Dig Dis Sci.* 2017;62(10):2840–6.
60. Herrero LA, Curvers WL, Bansal A, Wani S, Kara M, Schenk E, Schoon EJ, Lynch CR, Rastogi A, Pondugula K, et al. Zooming in on Barrett oesophagus using narrow-band imaging: an international observer agreement study. *Eur J Gastroenterol Hepatol.* 2009;21(9):1068–75.
61. Gross S, Trautwein C, Behrens A, Winograd R, Palm S, Lutz HH, Schirin-Sokhan R, Hecker H, Aach T, Tischendorf JJ. Computer-based classification of small colorectal polyps by using narrow-band imaging with optical magnification. *Gastrointest Endosc.* 2011;74(6):1354–9.
62. Hotta K, Kudo S, Mori Y, Ikematsu H, Saito Y, Ohtsuka K, Misawa M, Itoh H, Oda M, Mori K. Computer-aided diagnosis for small colorectal lesions: a multi-center validation 'endobrain study' designed to obtain regulatory approval. *Gastrointest Endosc.* 2019;89(6):AB76.
63. Rajan P, Canto M, Gorospe E, Almario A, Kage A, Winter C, Hager G, Wittenberg T, Münnzenmayer C. Automated diagnosis of Barrett's esophagus with endoscopic images. In: World Congress on medical physics and biomedical engineering, September 7–12, 2009, Munich, Germany, p. 2189–2192, Springer, 2009.
64. van der Putten J, Wildeboer R, de Groot J, van Sloun R, Struyvenberg M, van der Sommen F, Zinger S, Curvers W, Schoon E, Bergman J, et al. Deep learning biopsy marking of early neoplasia in barrett's esophagus by combining WLE and BLI modalities. In: 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019). Piscataway: IEEE; 2019. p. 1127–31.
65. van der Putten J, Struyvenberg M, de Groot J, Curvers W, Schoon E, Baldaque-Silva F, Bergman J, van der Sommen F, et al. Endoscopydriven pretraining for classification of dysplasia in barrett's esophagus with endoscopic narrow-band imaging zoom videos. *Appl Sci.* 2020;10(10):3407.
66. Struyvenberg MR, de Groot AJ, van der Putten J, van der Sommen F, Baldaque-Silva F, Omae M, Pouw R, Bisschops R, Vieth M, Schoon EJ, et al. A computer-assisted algorithm for narrow-band-imaging–based tissue characterization in Barrett's esophagus. *Gastrointest Endosc.* 2020;93:89–98.
67. Kiesslich R, Goetz M, Vieth M, Galle PR, Neurath MF. Confocal laser endomicroscopy. *Gastrointest Endosc Clin.* 2005;15(4):715–31.
68. Grisan E, Veronese E, Diamantis G, Trovato C, Crosta C, Battaglia G. Computer aided diagnosis of barrett's esophagus using confocal laser endomicroscopy: preliminary data. *Dig Liver Dis.* 2012;44:S147–8.
69. Veronese E, Grisan E, Diamantis G, Battaglia G, Crosta C, Trovato C. Hybrid patch-based and image-wide classification of confocal laser endomicroscopy images in barrett's esophagus surveillance. In: 2013 IEEE 10th international symposium on biomedical imaging. Piscataway: IEEE; 2013. p. 362–5.
70. Hong J, Park B-y, Park H. Convolutional neural network classifier for distinguishing barrett's esophagus and neoplasia endomicroscopy images. In: 2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC). Piscataway: IEEE; 2017. p. 2892–5.
71. Ghatwary N, Ahmed A, Ye X, Jalab H. Automatic grade classification of barretts esophagus through feature enhancement. In: Medical imaging 2017: computer-aided diagnosis, vol. 10134. International Society for Optics and Photonics, Washington; 2017. p. 1013433.
72. Ghatwary N, Ahmed A, Grisan E, Jalab H, Bidaut L, Ye X. In-vivo Barrett's esophagus digital pathology stage classification through feature enhancement of confocal laser endomicroscopy. *J Med Imaging.* 2019;6(1):014502.
73. Pulido JV, Guleria S, Ehsan L, Shah T, Syed S, Brown DE. Screening for Barrett's esophagus with probe-based

- confocal laser endomicroscopy videos. In: 2020 IEEE 17th international symposium on biomedical imaging (ISBI). Piscataway: IEEE; 2020. p. 1659–63.
74. Huang D, Swanson EA, Lin CP, Schuman JS, Stinson WG, Chang W, Hee MR, Flotte T, Gregory K, Puliafito CA, et al. Optical coherence tomography. *Science*. 1991;254(5035):1178–81.
75. Qi X, Sivak MV Jr, Wilson DL, Rollins AM. Computer-aided diagnosis of dysplasia in Barrett's esophagus using endoscopic optical coherence tomography. In: Coherence domain optical methods and optical coherence tomography in biomedicine VIII, vol. 5316. Bellingham: International Society for Optics and Photonics; 2004. p. 33–40.
76. Qi X, Rowland DY, Sivak MV Jr, Rollins AM. Computer-aided diagnosis of dysplasia in Barrett's esophagus using multiple endoscopic OCT images. In: Coherence domain optical methods and optical coherence tomography in biomedicine X, vol. 6079. Bellingham: International Society for Optics and Photonics; 2006. p. 60790I.
77. Rodriguez-Diaz E, Singh SK. Computer-assisted image interpretation of volumetric laser endomicroscopy in barrett's esophagus. *Gastroenterology*. 2015;148(4):S-91.
78. Scheeve T, Struyvenberg MR, Curvers WL, de Groot AJ, Schoon EJ, Bergman JJ, van der Sommen F, et al. A novel clinical gland feature for detection of early Barrett's neoplasia using volumetric laser endomicroscopy. In: Medical imaging 2019: computer-aided diagnosis, vol. 10950. International Society for Optics and Photonics, Washington; 2019. p. 109501Y.
79. van der Sommen F, Klomp SR, Swager A-F, Zinger S, Curvers WL, Bergman JJ, Schoon EJ, de With PH. Predictive features for early cancer detection in barrett's esophagus using volumetric laser endomicroscopy. *Comput Med Imaging Graph*. 2018;67:9–20.
80. van der Putten J, van der Sommen F, Struyvenberg M, de Groot J, Curvers W, Schoon E, Bergman JJ, et al. Tissue segmentation in volumetric laser endomicroscopy data using fusionnet and a domain-specific loss function. In: Medical imaging 2019: image processing, vol. 10949. International Society for Optics and Photonics, Washington; 2019. p. 109492J.
81. Fonollà R, Scheeve T, Struyvenberg MR, Curvers WL, de Groot AJ, van der Sommen F, Schoon EJ, Bergman JJ, et al. Ensemble of deep convolutional neural networks for classification of early barrett's neoplasia using volumetric laser endomicroscopy. *Appl Sci*. 2019;9(11):2183.
82. van der Putten J, Struyvenberg M, de Groot J, Scheeve T, Curvers W, Schoon E, Bergman JJ, de With PH, van der Sommen F. Deep principal dimension encoding for the classification of early neoplasia in barrett's esophagus with volumetric laser endomicroscopy. *Comput Med Imaging Graph*. 2020;80:101701.
83. Oh J, Hwang S, Tavanapong W, De Groot PC, Wong J. Blurryframe detection and shot segmentation in colonoscopy videos. In: Storage and retrieval methods and applications for multimedia, SPIE, Washington; 2004, vol. 5307; 2003. p. 531–42.
84. Hwang S, Oh J, Lee J, Cao Y, Tavanapong W, Liu D, Wong J, De Groot PC. Automatic measurement of quality metrics for colonoscopy videos. In: Proceedings of the 13th annual ACM international conference on multimedia. New York: Association for Computing Machinery; 2005. p. 912–21.
85. Wang Q, Pan N, Xiong W, Lu H, Li N, Zou X. Reduction of bubble-like frames using a rss filter in wireless capsule endoscopy video. *Opt Laser Technol*. 2019;110:152–7.
86. Pietri O, Rezgui G, Histace A, Camus M, Nion-Larmurier I, Li C, Becq A, Abou Ali E, Romain O, Chaput U, et al. Development and validation of an automated algorithm to evaluate the abundance of bubbles in small bowel capsule endoscopy. *Endosc Int Open*. 2018;6(4):E462.
87. Akbari M, Mohrekesh M, Najariani K, Karimi N, Samavi S, Soroushmehr SR. Adaptive specular reflection detection and inpainting in colonoscopy video frames. In: 2018 25th IEEE international conference on image processing (ICIP). Piscataway: IEEE; 2018. p. 3134–8.
88. Van Dongen N, van der Sommen F, Zinger S, Sekoon E, de With P. Automatic assessment of informative frames in endoscopic video. In: 2016 IEEE 13th international symposium on biomedical imaging (ISBI). Piscataway: IEEE; 2016. p. 119–22.
89. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Liang J. Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE Trans Med Imaging*. 2016;35(5):1299–312.
90. Islam AR, Alammari A, Oh J, Tavanapong W, Wong J, de Groot PC. Non-informative frame classification in colonoscopy videos using CNNs. In: Proceedings of the 2018 3rd international conference on biomedical imaging, signal processing; 2018. p. 53–60.
91. Hong D, Tavanapong W, Wong J, Oh J, De Groot PC. 3D reconstruction of virtual colon structures from colonoscopy images. *Comput Med Imaging Graph*. 2014;38(1):22–33.
92. van der Putten J, de Groot J, van der Sommen F, Struyvenberg M, Zinger S, Curvers W, Schoon E, Bergman J, de With PH. Informative frame classification of endoscopic videos using convolutional neural networks and hidden markov models. In: 2019 IEEE international conference on image processing (ICIP). IEEE, Taipei; 2019. p. 380–4.
93. Boers T, van der Putten J, de Groot J, Struyvenberg M, Fockens K, Curvers W, Schoon E, van der Sommen F, Bergman J, et al. Detection of frame informativeness in endoscopic videos using image quality and recurrent neural networks. In: Medical imaging 2020: image processing, vol. 11313. International Society for Optics and Photonics, Washington; 2020. p. 1131315.

94. Boers T, van der Putten J, Struyvenberg M, Fockens K, Jukema J, Schoon E, van der Sommen F, Bergman J, et al. Improving temporal stability and accuracy for endoscopic video tissue classification using recurrent neural networks. *Sensors.* 2020;20(15):4133.
95. Arribas J, Antonelli G, Frazzoni L, Fuccio L, Ebibgo A, van der Sommen F, Ghatwary N, Palm C, Coimbra M, Renna F, et al. Standalone performance of artificial intelligence for upper GI neoplasia: a meta-analysis. *Gut.* 2020;70:1458–1468.
96. Swager A-F, van der Sommen F, Klomp SR, Zinger S, Meijer SL, Schoon EJ, Bergman JJ, Peter H, Curvers WL. Computer-aided detection of early barrett's neoplasia using volumetric laser endomicroscopy. *Gastrointest Endosc.* 2017;86(5):839–46.



Artificial Intelligence for Colorectal Polyps in Colonoscopy

69

Luisa F. Sánchez-Peralta, J. Blas Pagador, and Francisco M. Sánchez-Margallo

Contents

Introduction	968
Components for Developing DL	970
Datasets	974
Polyp Detection and Localization	975
Polyp Segmentation	976
Polyp Classification	977
Conclusions and Future Trends	978
Cross-References	979
References	979

Abstract

Colorectal cancer (CRC) is one of the leading causes of death worldwide. Fortunately, its early detection and treatment highly improves the survival rates and reduces costs. In this regard, screening programs and colonoscopy

play an essential role. Artificial intelligence (AI) and deep learning (DL) have arisen with great success and been widely applied to medical imaging for the last few years and efforts have also been placed to develop methods to improve the adenoma detection rate in colonoscopy. In this chapter, polyp detection, localization, segmentation, and classification using colonoscopy are addressed. Works on these tasks have shown an exponential growth in the last few years. In the first place, the elements required for applying supervised DL methods for colorectal polyps in colonoscopy are introduced. The focus is placed on the model architecture, the dataset, the loss function, data augmentation, and metrics. Next, the currently openly available datasets that might be useful for future research are presented, before discussing methods for polyp detection, localization, segmentation, and classification.

Supplementary Information: The online version of this chapter (https://doi.org/10.1007/978-3-030-64573-1_308) contains supplementary material, which is available to authorized users.

L. F. Sánchez-Peralta (✉) · J. B. Pagador
Bioengineering and Health Technologies, Jesús Usón
Minimally Invasive Surgery Centre, Cáceres, Spain
e-mail: lf.sanchez@ccmijesususon.com;
jbpagador@ccmijesususon.com

F. M. Sánchez-Margallo
Scientific Direction, Jesús Usón Minimally Invasive
Surgery Centre, Cáceres, Spain
e-mail: msanchez@ccmijesususon.com

Lastly, some challenges and future trends in these fields are commented. This chapter about AI and DL for colorectal polyps in colonoscopy put forward the wide range of applications and research lines that can be further investigated with promising results as shown by the already currently published works, and always with the ultimate goal of assisting the endoscopist to improve CRC detection and diagnosis, which would eventually lead to better patient outcomes.

Keywords

Colorectal polyps · Detection · Localization · Segmentation · Classification · Colorectal cancer · Artificial intelligence · Deep learning

Introduction

Colorectal cancer (CRC) is defined as a carcinoma, most commonly an adenocarcinoma, located in the colon or rectum [1]. It is estimated that the number of incident cases and number of deaths will double from 2018 to 2040, reaching over one million deaths yearly [2], even though early detection of CRC highly improves the 5-year survival rate up to 88.5% if detected in an initial stage [3], which also implies a high reduction of costs [4]. In this regard, screening programs and colonoscopy play an essential role. Colonoscopy is the visual exploration of the colon and rectum by inserting a flexible endoscope through the anus [5]. During the procedure, precursor lesions of CRC can be detected, so the most convenient strategy can follow. This might range from leaving the lesion and scheduling a surgical intervention for its removal if it is an advanced stage, to the strategies of “diagnose and leave behind” or “resect and discard” of diminutive polyps that present a low risk of developing CRC, as long as their histopathology can be assessed in real time with high accuracy [6]. Nevertheless, the most common and traditional approach is to remove the lesion during colonoscopy and send it for pathological analysis. In any case, the removal of pre-malign lesions, if deemed

necessary, is associated with the reduction of CRC mortality [7]. The adenoma detection rate (ADR) of the endoscopist is defined as the percentage of colonoscopies performed with at least one adenoma identified. It is shown that the higher the ADR is, the lower the rate of detecting CRC after screening examination [8]. This is also known as interval CRC and that can be defined as a CRC “diagnosed after a screening or surveillance examination in which no cancer is detected, and before the date of the next recommended examination” [9].

Computer-Assisted Detection (CADe) and Computer-Assisted Diagnosis (CADx) have the potential to improve colonoscopy on three key areas: (1) adequacy of mucosal inspection, (2) polyp detection, and (3) optical biopsy [10]. This chapter is focused on the last two. Moreover, artificial intelligence (AI) and deep learning (DL) have arisen with great success and been widely applied to medical imaging for the last few years, obtaining a similar performance to experts in many fields. Undoubtedly, efforts have also been placed to develop AI and DL methods to improve the ADR in colonoscopy, although more randomized controlled trials are still necessary [11]. These methods can be designed to fulfill one or more of the following tasks related to CADe and CADx for colorectal polyps (Fig. 1):

1. Polyp detection: In this case, the aim is to label the frame, indicating whether a polyp is shown or not in it, without further information. In this chapter, image and frame will be used indistinctly, as images are actually frames from a video or can be considered as such.
2. Polyp localization: In this task, the position of the polyp is given within the corresponding frame, either by a bounding-box, a circle, or the center coordinates.
3. Polyp segmentation: In the frame, a precise delimitation of the polyp area is obtained.
4. Polyp classification: For a given polyp, the objective is to label it as benign/malign or as a category of any classification schema, such as Paris [12], Kudo [13] or NICE [14]. This can be considered as “optical biopsy,” as it allows for an *in situ* diagnosis.

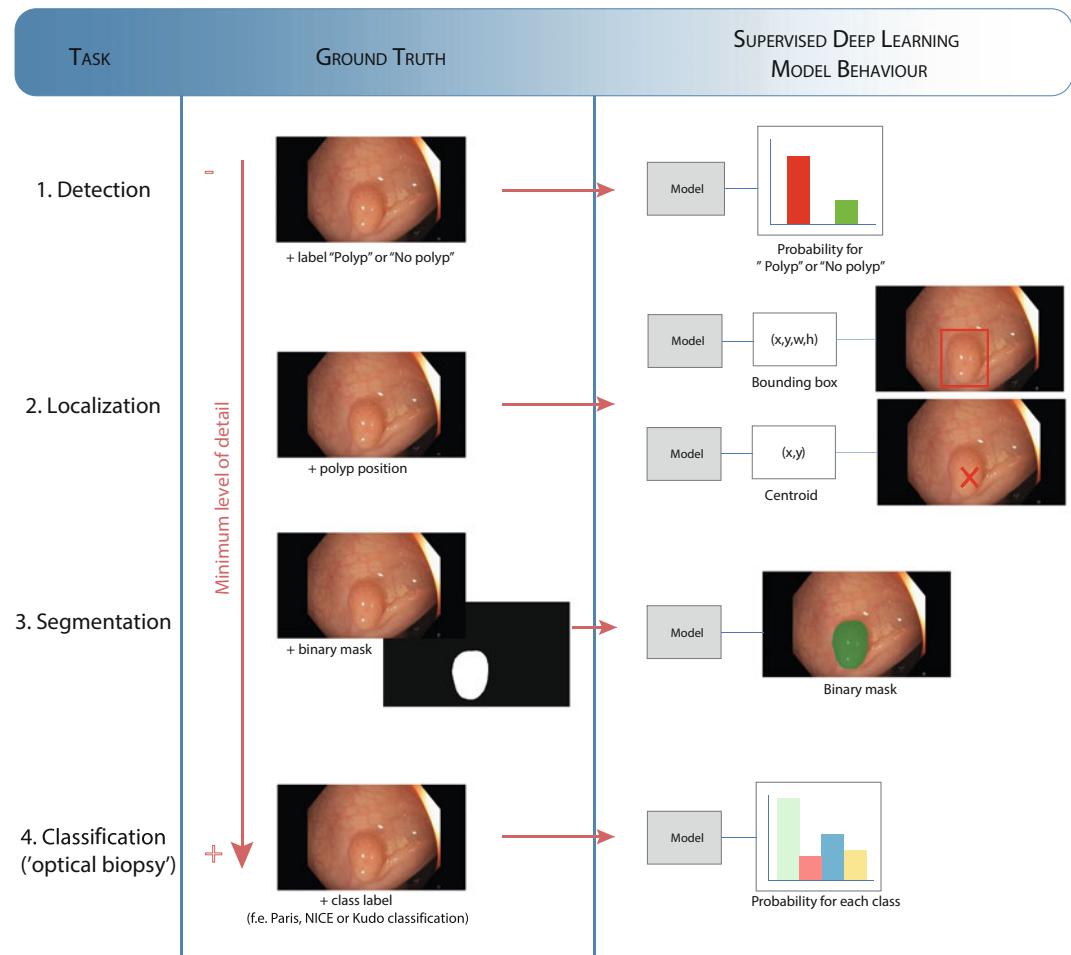


Fig. 1 Tasks considered in this chapter. Ground truth and model behavior for the four tasks are also included

Several recent reviews gather the wide spectrum of methods for detection, localization, segmentation, and classification of colorectal polyps [15–18]. These fields have had great success for the last few years. A quick search only on the Web of Science using the search string “deep learning” AND “colorectal” AND “polyp” AND (detection OR localization OR segmentation OR classification) shows the exponential growth in the field of this chapter (Fig. 2). It is clear that the increasing technical capacities, together with the high impact on mortality that early detection of CRC has, are essential to keep moving forward in this field. Currently, some of the published methods have already been deployed into the market, such as GI Genius module from Medtronic, CAD EYE from

Fujifilm, or AI4G from AI4GI and Olympus [19]. Lastly, it is also worth mentioning the value of AI for the diagnose-and-leave strategy, as it saves cost of unnecessary polypectomies and pathologic examinations, meaning an estimated reduction between 6.9% and 18.9% of the gross annual reimbursement for colonoscopies upon the country [20].

This chapter is organized as follows: Firstly, the elements required for applying supervised DL methods for colorectal polyps in colonoscopy are described. The currently openly available datasets that might be useful for future research are presented, to then further discuss methods for polyp detection, localization, segmentation, and classification. For the sake of the chapter extension,

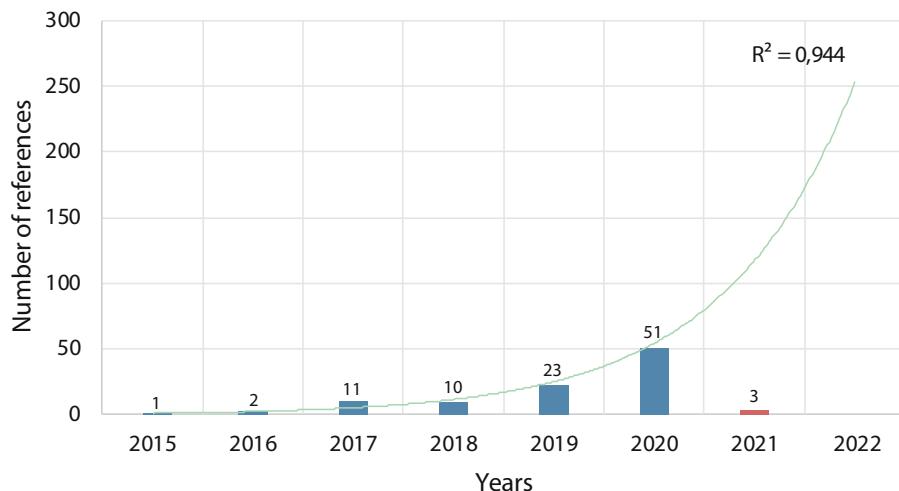


Fig. 2 Trend for publications in the fields of detection, localization, segmentation, and classification methods for colorectal polyps using deep learning (search performed on 25 January 2021)

detailed references to reviews are included, and the reader is encouraged to peruse them for further description of the AI and DL methods. Lastly, we comment on some challenges and future trends in these fields.

Components for Developing DL

Before going into detail, the components that should be considered when developing a DL model for colorectal polyps are briefly revised (Fig. 3):

1. An annotated dataset of colorectal polyps is needed. To this end, a proprietary dataset or any of the openly available datasets presented in the following section can be used. As indicated in Fig. 1, ground truth information might have different level of detail that should be directly related to the desired task to accomplish, but in any case, it should be provided by a clinical expert. For instance, at least the position of the polyp would be necessary for its localization; as labels would not be enough. It is also possible to take advantage of larger datasets not related to colorectal polyps, such as ImageNET, MSCoco, or PascalVOC, by using them to pretrain either the encoder/

classification network or the whole model, which will be then fine-tuned with the dataset of colorectal polyps. When splitting the dataset into training, validation, and test set, patient independence between sets must be assured. Each set is intended to be used at a different point during the process.

2. Strategies to overcome scarce and/or weak annotations might also be needed [21]. Medical imaging datasets are usually limited in size, so one of the most common strategies is data augmentation that consist of an artificial increase of the training set by transforming the original image. Among the possible transformations, they can be classified depending on the level at which they are applied into [22]:
 - (a) Image-level: such as width and height shifts, rotation, or shear. In these cases, the image is modified, and while this does not alter the label assigned to the image, it requires that the binary mask, if available, is modified in the same way.
 - (b) Pixel-level: such as changes in brightness and contrast. These modifications do not alter the label or the mask, as they only affect the pixel values of the RGB channels.
 - (c) Problem-based: these transformations mimic particularities of the endoscopic

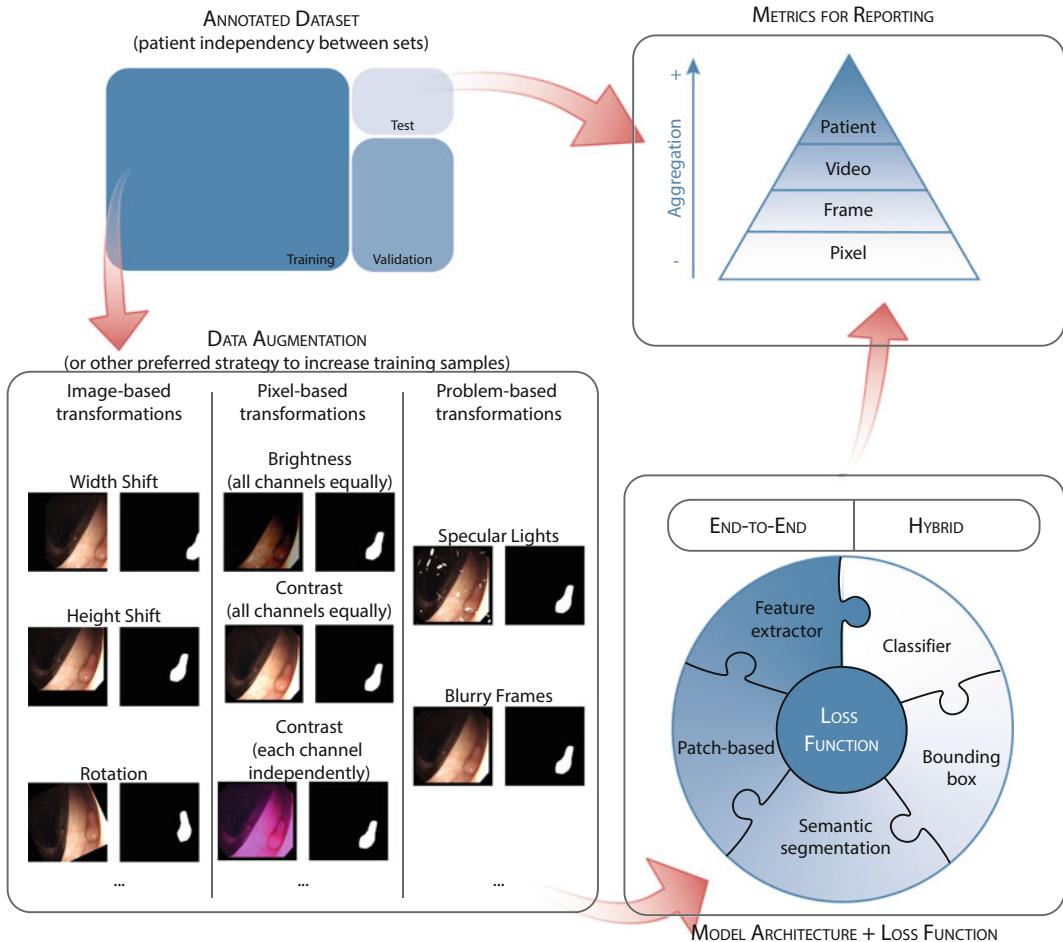


Fig. 3 Schematic setup of DL methods for colorectal polyps and necessary elements

images, such as the presence of specular lights or blurry frames.

3. An appropriate model architecture matching our objective task should be selected, created, and/or modified. In a previous work, a two-level classification scheme was defined [15]. In the first level, methods are classified into end-to-end, if one single DL approach is used to accomplish the task, and hybrid methods, if a DL approach is combined with other hand-crafted method. On the second level, we defined five approaches depending on their use of the DL architecture (Fig. 4):
 - (a) Feature extractor: The model is used to automatically create a feature vector that

is the input of a traditional classifier, such as Support Vector Machine.

- (b) Classification: A classification network is used in detection tasks to label an image as containing a polyp or not, without positioning of the polyp, or in the classification task to assign the most probable class of the classification scheme.
- (c) Patch-based: The presence of polyp is obtained for each image patch or tile, so the location of the polyp might be obtained based on the patch location.
- (d) Bounding-box: The location of the polyp through a bounding-box (coordinates of the upper right corner, height, and width)

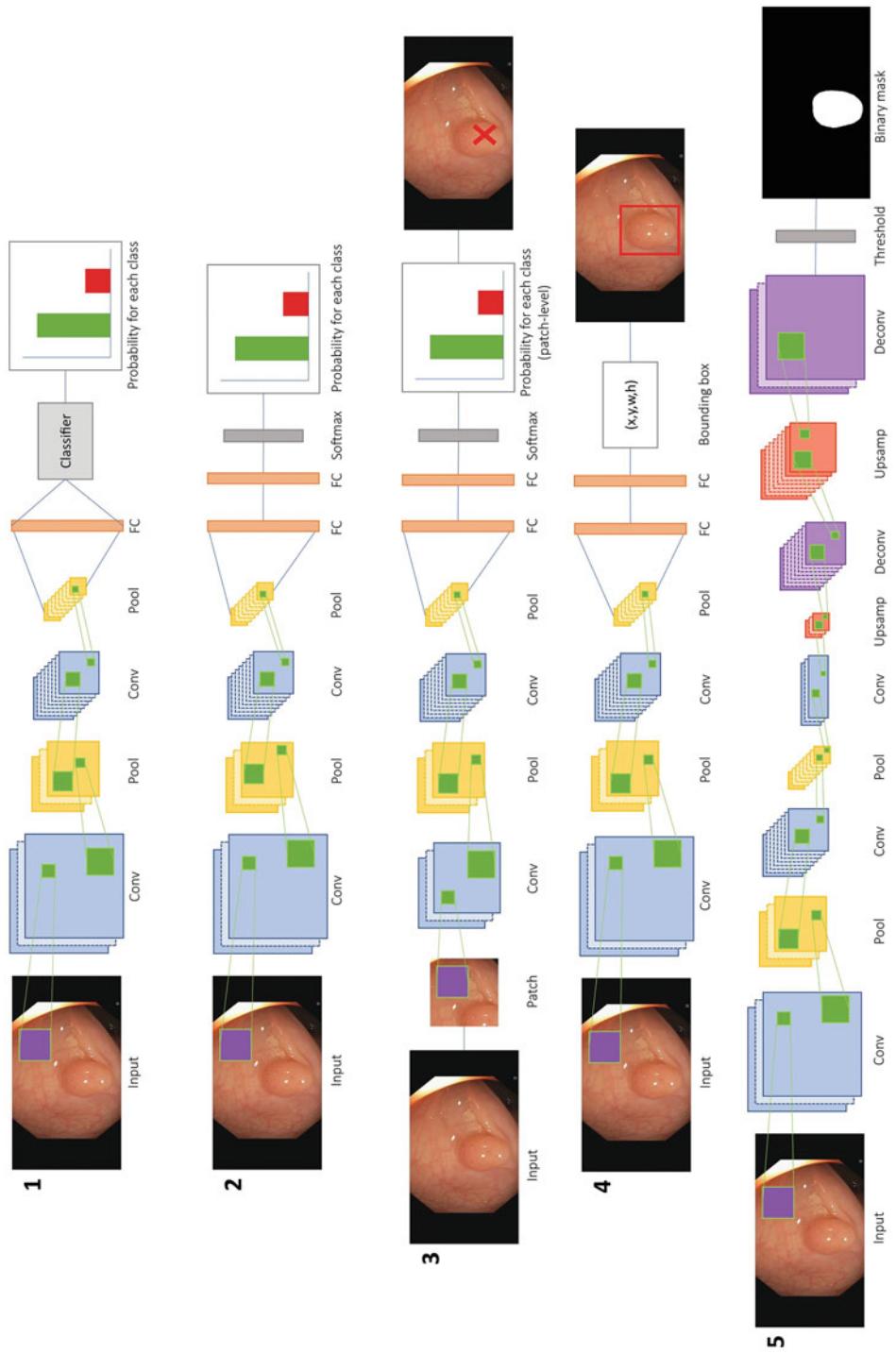


Fig. 4 Schematic representation of different approaches for polyp detection, localization, segmentation, and/or classification. From top to bottom, (1) feature extractor, (2) classification, (3) patch-based, (4) bounding-box, and (5) semantic segmentation. Each type of layer is represented by a different color: convolutional layer (conv), pooling layer (pool), fully connected layer (FC), upsampling layer (upsamp), and deconvolutional layer (deconv). The receptive field is marked with a green square. (Reproduced from [15] under Creative Commons License CC BY-NC-ND 4.0)

Fig. 4 Schematic representation of different approaches for polyp detection, localization, segmentation, and/or classification. From top to bottom, (1) feature extractor, (2) classification, (3) patch-based, (4) bounding-box, and (5) semantic segmentation. Each type of layer is represented by a different color: convolutional layer (conv),

is given by the DL method, generally using a regression layer.

- (e) Semantic segmentation: The image is labeled at a pixel-level either as polyp or background. Encoder-decoder networks are usually selected for this task. While encoder or first half of the layers encode the image description highlighting the discriminative features, the decoder or second half is responsible for mapping the low-resolution encoding into full input resolution feature maps.

End-to-end methods are surpassing hybrid ones, thanks to the increasing computing capacity of systems. Convolutional Neural Networks (CNNs) are the most used DL architectures for detection, localization, and classification, while Fully Convolutional Networks (FCNs) are preferred for segmentation tasks [15, 18]. Other architectures, such as Recurrent Neural Networks (RNNs) or Generative Adversarial Networks (GANs), can also be found in the literature for some of the considered tasks in this chapter. Classical architectures as AlexNet, VGG16, or VGG19, and GoogLeNet are widely used as backbone/feature extractor/encoder in our field. These FCNs can be initialized from the weights of a classification network (encoder) that acts as a feature extractor.

4. A suitable loss function should be employed, bearing in mind the target task. In this regard, we can use either widely known loss functions, such as binary cross entropy and Jaccard loss function for classification and segmentation, respectively, or custom-made loss functions [23]. When selecting the loss function, it is advisable to take into account the distribution of the dataset, i.e., the distribution of polyp/no polyp images, the proportion of pixels labeled as polyp or no polyp, or the proportion of examples for each class in a classification task. With balanced datasets, a traditional loss function might suffer; otherwise, a specific loss function might be necessary to overcome the imbalance toward the positive class.

5. Selection of reporting metrics, which will compare the obtained result of the model with the ground truth in any format (label, bounding-box, and/or binary mask). In most cases, they are based on the confusion matrix: true positives (TN), true negative (TN), false positives (FP), and false negatives (FN), such as accuracy, precision, recall, and specificity as well as combination of these, such as F1-score and F2-score. Furthermore, widely used overlap measures, such as the Jaccard index or Dice index, can also be determined based on the confusion matrix elements. It is highly recommended to report results in more than one single metric [18]. Although this issue might seem irrelevant, a poor selection of metrics might result on artificially increasing the DL performance and failing to show if the model works correctly and in the best way. For polyp detection, localization, segmentation, and detection, several reviews [15, 16, 18] identified the wide range of metrics used, which makes more difficult to fairly compare methods. Furthermore, the same metric can be calculated at different levels: pixel, frame, video, or patient-level, for a different level of aggregation. Depending on the task, one or another might be more convenient. Figure 5 illustrates different casuistries for a segmentation task, where six frames from three polyps compose the test set. While the accuracy, which includes TN in the numerator, reaches values above 0.750 in all cases, the Jaccard index, as overlap measure, penalizes both FP and FN, showing differences between the different frames A to F. The analysis of metrics per polyp instead of considering all frames together, regardless of their origin, lead also to different results. In this example, polyp 3 is much worse segmented than polyp 1 in average, but this cannot be identified if the test set is analyzed as a whole; therefore a per polyp analysis could give further information on which ones are poorly detected and/or segmented to further improve the DL method in those particular cases. Similarly, and assuming detection if the Jaccard index is above the 0.5 threshold, differences can be also observed

		POLYP 1			POLYP 2			POLYP 3			
		Frame A	Frame B	Frame C	Frame D	Frame E	Frame F				
TP	20	15	17	17	5	5	5	Average			
TN	80	75	78	58	80	70					
FP	0	5	5	10	0	15					
FN	0	5	0	15	15	10					
Accuracy	1.000	0.900	0.950	0.750	0.850	0.750		0.867			
per polyp	0.950			0.800				0.833			
Jaccard	1.000	0.600	0.773	0.405	0.250	0.167		0.532			
per polyp	0.791			0.208				0.468			
detection ($J > 0.5$)	1	1	1	0	0	0		0.500			
per polyp	1			0				0.333			

Fig. 5 Comparison of metrics. TP: True Positives. TN: True Negatives. FP: False Positives. FN: False Negatives. Accuracy is calculated as $(TP+TN)/(TP+TN+FP+FN)$, while

Jaccard is calculated as $TP/(TP+FP+FN)$. Dashed line indicates the polyp ground truth

between the analysis per polyp or per frame, decreasing the capacity of detection from 50% to 33%.

It is also important to highlight that some features of the endoscopic images might have an impact on the image processing analysis [17], which is the basis of any AI or DL method. On one side, features derived from the acquisition process, such as color artifacts or interlacing and, more importantly, the presence of specular lights due to the perpendicular incidence of the endoscope light source onto a shiny surface. On the other side, there are also features derived from the endoscopic image per se, such as a wide variety of polyp appearances.

Datasets

The limited size of datasets for medical imaging analysis using AI and DL is one well-known challenge to address. Having to annotate and/or segment the images, videos, or studies is a cumbersome, time-consuming task. Furthermore, this process must be done by clinical experts with

wide knowledge and experience, who usually lack time for these tasks. Therefore, there is still a need to create larger, publicly available datasets in a collaborative manner. Nevertheless, there are currently different public datasets that can be used to train DL models in the field of colorectal polyps in colonoscopy:

1. CVC-EndoSceneStill [24] is a manually segmented dataset compiling and improving two other datasets, CVC-ColonDB and CVC-ClinicDB. It includes 912 images from 44 video sequences from 36 patients. Annotations include a binary mask for the lumen, specular lights, polyp, and void areas. Furthermore, division into training, validation, and test sets is provided by the owners.
2. CVC-VideoClinicDB [25, 26] gathers 18 video sequences showing a polyp. In this case, annotations are binary masks showing ellipses that approximate the polyp area. This way, annotated frames come to 9.221 polyp frames, out of a total of 10.924 frames.
3. ETIS-Larib [27, 28] provides 196 white light (WL) images from 34 sequences showing

- 44 different polyps. The provided ground truth is binary masks manually annotated.
4. ASU-Mayo Clinic [29] provides binary masks for 20 annotated videos as well as other 18 videos for testing for which ground truth is not available. Frames include both WL and narrow band imaging (NBI).
 5. Kvasir-SEG [30] provides a manually delimitated binary mask for 1.000 polyp images, together with the corresponding bounding-boxes in a JSON file. It is also part of a larger dataset for gastrointestinal endoscopy called HyperKvasir [31].
 6. PICCOLO WL and NBI colonoscopic dataset [32] comprises 3.433 polyp frames from 76 lesions from 40 patients, also divided into training, validation, and test sets assuring patient independence between them. This dataset also provides clinical metadata of the lesions, including size, Paris and NICE classification, and histological diagnosis.
 7. Mesejo et al. [33] provide WL and NBI videos for 15 serrated adenomas, 21 hyperplastic lesions, and 40 adenoma.

Furthermore, Nogueira-Rodríguez et al. [16] provide an extensive summary of proprietary datasets, which have been used by several authors when they had access to such clinical information.

Depending on the aimed task, one or another dataset might be more convenient, depending on the ground truth provided. In any case, they might also be used in conjunction, both public and proprietary, to increase the number and variety of samples, either for training or to assure transferability at testing phase. In this regard, three public datasets, CVC-EndoSceneStill, Kvasir-SEG, and PICCOLO datasets have been compared, finding that PICCOLO shows better generalization capabilities [32].

The datasets should gather all range of polyp variability. While this is the ideal situation, it is still far from being accomplished since flat polyps are usually underrepresented in those public datasets. Flat polyps are the most difficult ones to detect [32], so CADe systems would play an essential role to assist endoscopists to reduce their miss rate. A joint global initiative would help

(1) to collect a larger sample of abnormal cases like these flat polyps, which are usually less frequent; (2) to have good quality annotations; and (3) to increase variability of the acquisition devices, which would eventually result in more robust AI and DL methods.

One last remark is the need for a common reporting framework. In order to perform a fair comparison of the different methods, the same images and metrics should be used. Therefore, it would be advisable to establish a set of criteria and guidelines for the authors to follow. Patient independency between training and reporting sets must be assured in any case, so it is highly encouraged that the dataset owners establish a distribution into training, validation, and test sets or the definition of a bootstrapping methodology.

Polyp Detection and Localization

Detecting and locating a polyp is the trigger event for CRC diagnosis. Therefore, this is directly related to improving the ADR, as DL models will help endoscopists to ensure that no lesions are missed during the colonoscopy procedure that would eventually develop CRC. Therefore, these DL methods have a highly important clinical implication and the eventual increment of the ADR, will lead to a reduction of interval CRC and its associated mortality. In this regard, subtle lesions are prone to be missed even for trained endoscopists, so CADe systems would play an important role. In this regard, Guo et al. [34] have analyzed a YOLOv3 CADe system, which achieved a similar sensitivity to experts and superior to physicians in training, using 50 short videos showing one or two polyps of mean size 3.5 ± 1.5 mm. This fact emphasizes the utility of AI and DL for diminutive, subtle lesions that might be missed even by more experienced endoscopists.

There are several reviews that summarize the state of the art of detection and localization methods either from a technical or a clinical point of view. From a more clinical approach, Barua et al. [35] analyzed five randomized trials and concluded that the number of small adenomas

(≤ 5 mm) found in colonoscopy assisted by AI methods was higher than without AI assistance, while the number of large adenomas (> 5 mm) remained similar. On the other hand, placing the focus on the technical aspects of DL methods, Sanchez-Peralta et al. [15] systematically reviewed the state of the art and found 26 works related to detection and localization. They found that end-to-end and hybrid methods have an equal presence, but the trend is toward the use of end-to-end methods. In a similar approach, Nogueira-Rodríguez et al. [16] included 21 works for detection and localization in conventional colonoscopy, identifying those able to detect multiple polyps in real time. In a more extensive manner, Pacal et al. [18] reviewed polyp detection and localization in different imaging modalities, including 35 works that use colonoscopy.

Nevertheless, most papers gathered in those presented reviews are focused on still frames from the datasets, either public or proprietary. In this situation, the temporal coherence of the presence of a polyp in a video cannot be exploited. Since the polyp would not abruptly disappear from one frame to the next one, the use of videos instead of still frames would allow the inclusion of this temporal dependency. In this line, a recent work by González-Bueno Puyal et al. [36] proposed a 2D/3D CNN for segmentation where five consecutive frames are considered as compromised between new temporal information and detection speed. Another relevant aspect is the reduction of frequent false positives since they may lead to operator burnout and/or diminished trust in the system. Qadir et al. [37] proposed a two-stage method that combines a CNN-based detector with a false positive reduction unit to overcome this situation. In this regard, Holzwanger et al. [38] recommended a time-based threshold of 2 s to maximize specificity and accuracy.

Regarding metrics, analysis of results for polyp detection and localization is usually done at image or frame-level, as either the presence or absence of polyps is indicated individually. Metrics based on the confusion matrix such as accuracy, precision, or recall are useful as they show the proportion of correct or incorrect labeled frames, and although

they are all widely used, recall has been found in a larger number of works [15]. Nevertheless, it would also be advisable to report detection per-polyp, establishing a threshold, such as Misawa et al. [39], who established that a polyp is correctly detected if the model identifies it more than 75% of the time it appears in the video. In the case of localization, and as long as a bounding-box is available as ground truth, overlap measures could be used, similarly to the task of polyp segmentation.

Polyp Segmentation

Once the presence of a polyp is confirmed in a given frame, it might be useful to determine the area that is considered to be lesion, for assisting on its removal if necessary. If this is done, a safety margin should also be taken into account. Therefore, for this task, segmentation methods should be applied to determine the precise contour of the polyp. Endoscopists will benefit from a precise polyp segmentation to assess the resection margins. In this case, it is essential to remove completely the lesion without leaving behind any part of it that could reproduce or evolve into CRC.

Similarly to the methods for detection and localization, segmentation models have also been reviewed in 2 works [15, 18] but their presence is not so wide as the former, finding 10 and 14 works, respectively. This might be due to the lack of widely annotated datasets, as it is less time consuming to label one frame than manually annotate the polyp area in an image. To overcome this, different solutions are provided for medical image in general [21]. Out of those, data augmentation is one of the most common techniques addressed for polyp segmentation [15, 16], although there is no consensus on the most beneficial transformations and they are mainly selected on a trial and error basis, and based on the researcher experience. In this regard, the effect of different transformations on two publicly available datasets has been studied [22] and it has been determined that, while polyp segmentation using CVC-EndoSceneStill benefits from pixel-based

transformations such as changes in brightness and contrast (even if they are not commonly used), if Kvasir-SEG is employed, then image-based transformations are recommended, especially rotation and shear. In any case, it is important to remark that including more intense data augmentation does not always lead to an increased performance [40].

For polyp segmentation, the combination of end-to-end and semantic segmentation models are the most commonly used methodology [15]; this means that the task is accomplished as a whole. Therefore, the use of encoder-decoder architectures is preferred. While the encoder transforms the input image into a feature vector capturing the context information, the decoder recovers the spatial information lost in the previous process. These encoder-decoder architectures can be based on off-the-self networks or designed on purpose for polyp segmentation. As an example of the first case, four different models are created by combining two backbones or encoders (VGG-16 [41] and Densenet121 [42]) with two different ways of joining the encoder and decoder (U-Net-based [43], so information is either concatenated, or added, if it is LinkNet-based [44]) to analyze the influence of the loss function on polyp segmentation [23]. On the other hand, Mahmud et al. [45] have designed the PolySegNet. Although it is based on U-Net, it incorporates three major building blocks to overcome initial limitations for polyp segmentation, so it achieves a Dice score over 84% in all the four public datasets used.

For polyp segmentation, overlap measures are generally taken into account. Either the Jaccard index, also known as Intersection over Union (IoU), or the Dice index are widely used, but despite this, it would also be advisable to further include distance metrics valid for small segments, such as Hausdorff distance or Mahalanobis distance [15]. Since semantic segmentation can also be seen as detection at pixel-level, by labeling each pixel as either belonging to the polyp or no polyp class, the metrics indicated in the polyp detection section might also be applied as long as they are calculated at pixel-level. In this case, it is important to emphasize that due to the

unbalanced situation, as the polyp area is generally much smaller than the background class, metrics including true negatives, such as the specificity, might obtain high values even when the polyp is poorly or not detected at all.

Polyp Classification

Lastly, classification methods allow for polyp characterization or an “optical biopsy” or “optical diagnosis,” that is, giving a clinical diagnosis in real time without the need of removing the lesion and sending the sample for histological analysis [46]. This way, patient safety might be increasing by not removing lesions without malignancy capacity, while time and costs are reduced. The strategy of “diagnose and leave behind” would benefit from real-time, in-situ accurate diagnosis, as long as the methods surpass the minimum value of 90% for the negative predictive value (NPV) requested by the American Society for Gastrointestinal Endoscopy [6].

Although polyp detection can also be seen as a classification between healthy tissue and lesion, in this section we refer to the determination of the type of lesion once it is detected. Reviews [16, 18] consistently found much less works for polyp classification based on WL images from colonoscopy, which possibly is due to the lack of clinical information in most of the publicly available datasets. This situation hinders the wider development of CADx systems by the scientific community and puts forward again the need for joint collaboration to create such datasets. Besides, this lack of methods is also due to the use of other imaging modalities, such as NBI, confocal endomicroscopy, or magnifying chromoendoscopy, for polyp characterization [46, 47].

Therefore, these classification methods might use any of the different classification approaches for colorectal lesions, but in most cases, they rely on NBI image rather than WL images [16], since in the clinical setting, WL is normally used for detection and once the lesion is detected, the imaging modality is switched to NBI, when available, for diagnosis, as this image modality highlights patterns useful for diagnosis. For example,

Patino-Barrientos et al. [48] used a pretrained VGG model as feature extractor to classify polyps according to the Kudo's pit pattern schema. They obtained 83% accuracy and F1-score, and their method outperformed traditional techniques where features are manually extracted. On a different scale, Rodriguez-Diaz et al. [49] classified previously segmented lesions into neoplastic (including tubular adenomas, tubulovillous adenomas, and adenocarcinomas) or non-neoplastic (including hyperplastic polyps and polypoid-appearing normal colonic mucosa). In this case, the NPV equaled 0.91 but the model is limited to near-focus NBI polyp images. More recently, Jin et al. [50] go a step further and included interpretable explanation to the optical diagnosis by overlapping a heatmap showing the probability within the image.

Metrics for polyp classification are usually calculated at frame level, although it would be more convenient to establish a diagnosis per polyp such as Byrne et al. [51], who trained a CNN to classify NBI frames into type 1 and 2 of the NICE classification, achieving an accuracy of 94% in 106 diminutive polyps.

Conclusions and Future Trends

Undoubtedly, AI and DL have come also to improve colonoscopy and polyp detection, localization, segmentation, and classification. The large amount of works over the last few years show that AI and DL methods have the potential to improve CADe and CADx systems in order to increase CRC early diagnosis and eventually reduce its associated mortality.

This potential is well recognized, and the American Society for Gastrointestinal Endoscopy has recently published a White Paper, where polyp detection and diagnosis are included as priority clinical use cases for developing AI algorithms [52]. There are also key research questions that remain unsolved. Ahmad et al. [53] followed a Delphi approach and determined guidance for future research on performance metrics, clinical trials design, and end points; technological developments,

clinical adoption, and integration into endoscopy; data (access, sharing/privacy, curation, and annotation); and regulatory approvals. Among these questions, it is worth to highlight the need to improve performance of more challenging and advanced lesions, such as subtle, flat lesions and sessile serrated lesions, as well as the need to reduce false positive rates to avoid the “alert fatigue,” and the need for real-time systems with minimal latency. Similarly, Hoerter et al. [54] also identified current challenges in terms of standardization of outcomes, training and testing datasets, real-world application, as well as regulatory approval and reimbursement. In all cases, it is clear that further studies are still necessary until a wide application of AI and DL methods for colorectal polyps in the clinical daily practice of colonoscopy. Furthermore, prospective clinical trials should be further carried out, in order to determine the actual impact of CADe and CADx systems in the ADR.

From a more technical point of view, this chapter has focused on supervised DL methods, and most of them include CNNs in their approaches. Therefore, there is still room for further research to apply more recent developments, initially applied on natural images. In this regard, semi-supervised or even unsupervised approaches could be considered. In these cases, the need for annotating images by experts would be reduced, so the creation of larger datasets would be easier. Up to now, there have been no efforts on semi-supervised methods for detection, localization, and segmentation [55]. For polyp classification, just recently Golhar et al. [56] applied a semi-supervised learning to improve classification into neoplastic/cancerous and non-neoplastic, which overcome the fully supervised approach, especially when few data are used.

Also, the use of Generative Adversarial Networks (GANs) could be further exploited. Its raise in natural images will for sure quickly evolve into a wider application for colorectal polyps in colonoscopy. So far, GANs are used for detection and localization [57], as well as for data augmentation [58], but the presence of these networks is

practically inexistent in the current reviews [15, 16, 18]. Doubtlessly this situation will change in the upcoming years, as other medical imaging modalities are already benefitting from the capabilities of GANs.

Furthermore, the use of synthetically created images to overcome the limited size of datasets could be analyzed. Initial efforts have already been done by Shin et al. [58] to create polyp images from normal colonoscopy images, although the ranges of color and texture of the created polyps is limited, or by De Almeida Thomaz et al. [59] to “copy and paste” polyps from one image to another in a convenient location. In any case, these efforts lead to determined polyps, and thus, the potential of synthetic images could still be further exploited to create non-deterministic polyps and cover the full range of clinical variability.

Therefore, this chapter about AI and DL for colorectal polyps in colonoscopy puts forward the wide range of applications and research lines that can be further investigated with promising results as shown by the already currently published works, and always with the ultimate goal of assisting the endoscopists to improve CRC detection and diagnosis, which would eventually lead to better patient outcomes.

Cross-References

- [AIM in Endoscopy Procedures](#)
- [AIM in Oncology](#)
- [Artificial Intelligence in Gastroenterology](#)
- [Basic Concepts of Artificial Intelligence: Primed for Clinicians](#)

References

1. World Health Organization. World cancer report 2014. 2014.
2. International Agency for Research on Cancer. Cancer tomorrow. 2020. <https://gco.iarc.fr/tomorrow/home>. Accessed 30 Nov 2020.
3. Wiegering A, Ackermann S, Riegel J, et al. Improved survival of patients with colon cancer detected by screening colonoscopy. *Int J Colorectal Dis.* 2016;31: 1039–45. <https://doi.org/10.1007/s00384-015-2501-6>.
4. Mar J, Errasti J, Soto-Gordoa M, et al. The cost of colorectal cancer according to the TNM stage. *Cirugía Española* (English Ed). 2017;95:89–96. <https://doi.org/10.1016/j.cireng.2017.01.001>.
5. Williams CB. Insertion technique. In: Waye JD, Rex DK, Williams CB, editors. *Colonoscopy. Principles and practice*. Blackwell Publishing; 2005.
6. Ferlitsch M, Moss A, Hassan C, et al. Colorectal polypectomy and endoscopic mucosal resection (EMR): European Society of Gastrointestinal Endoscopy (ESGE) Clinical Guideline. *Endoscopy*. 2017;49:270–97.
7. Berros Fombella JP, Aguilar Huergo S, García Teijido P. Enfermedades premalignas. In: Sociedad Española de Oncología Médica (ed) *Manual SEOM de prevención y diagnóstico precoz del cáncer*. 2017.
8. Lund M, Trads M, Njor SH, et al. Quality indicators for screening colonoscopy and colonoscopist performance and the subsequent risk of interval colorectal cancer: a systematic review. *JBI Database Syst Rev Implement Reports Online Fir*. 2019.
9. Ertem FU, Ladabaum U, Mehrotra A, et al. Incidence of interval colorectal cancer attributable to an endoscopist in clinical practice. *Gastrointest Endosc.* 2019;88:705–11. <https://doi.org/10.1016/j.gie.2018.05.012.Incidence>.
10. Byrne MF, Shahidi N, Rex DK. Will computer-aided detection and diagnosis revolutionize colonoscopy? *Gastroenterology*. 2017;153:1460–1464.e1. <https://doi.org/10.1053/j.gastro.2017.10.026>.
11. Aziz M, Fatima R, Dong C, et al. The impact of deep convolutional neural network-based artificial intelligence on colonoscopy outcomes: a systematic review with meta-analysis. *J Gastroenterol Hepatol*. 2020;1–8. <https://doi.org/10.1111/jgh.15070>.
12. Endoscopic Classification Review Group. Update on the Paris classification of superficial neoplastic lesions in the digestive tract. *Endoscopy*. 2005;37:570–8. <https://doi.org/10.1055/s-2005-861352>.
13. Kudo SE, Tamura S, Nakajima T, et al. Diagnosis of colorectal tumorous lesions by magnifying endoscopy. *Gastrointest Endosc.* 1996;44:8–14. [https://doi.org/10.1016/S0016-5107\(96\)70222-5](https://doi.org/10.1016/S0016-5107(96)70222-5).
14. Hayashi N, Tanaka S, Hewett DG, et al. Endoscopic prediction of deep submucosal invasive carcinoma: validation of the Narrow-Band Imaging International Colorectal Endoscopic (NICE) classification. *Gastrointest Endosc.* 2013;78:625–32. <https://doi.org/10.1016/j.gie.2013.04.185>.
15. Sánchez-Peralta LF, Bote-Curiel L, Picón A, et al. Deep learning to find colorectal polyps in colonoscopy: a systematic literature review. *Artif Intell Med.* 2020;108. <https://doi.org/10.1016/j.artmed.2020.101923>.
16. Nogueira-Rodríguez A, Domínguez-Carbajales R, López-Fernández H, et al. Deep neural networks approaches for detecting and classifying colorectal polyps. *Neurocomputing*. 2021;423:723–34. <https://doi.org/10.1016/j.neucom.2020.02.123>.
17. Sánchez-Montes C, Bernal J, García-Rodríguez A, et al. Review of computational methods for the

- detection and classification of polyps in colonoscopy imaging. *Gastroenterol Hepatol (N Y)*. 2020;43:222–32.
- 18. Pacal I, Karaboga D, Basturk A, et al. A comprehensive review of deep learning in colon cancer. *Comput Biol Med*. 2020;126:104003. <https://doi.org/10.1016/j.combiomed.2020.104003>.
 - 19. Wittenberg T, Raithel M. Artificial intelligence-based polyp detection in colonoscopy: where have we been, where do we stand, and where are we headed? *Visc Med*. 2020. <https://doi.org/10.1159/000512438>.
 - 20. Mori Y, Kudo S ei, East JE, et al. Cost savings in colonoscopy with artificial intelligence-aided polyp diagnosis: an add-on analysis of a clinical trial (with video). *Gastrointest Endosc*. 2020;92:905–11.e1. <https://doi.org/10.1016/j.gie.2020.03.3759>.
 - 21. Tajbakhsh N, Jeyaseelan L, Li Q, et al. Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation. *Med Image Anal*. 2020;63:101693. <https://doi.org/10.1016/j.media.2020.101693>.
 - 22. Sánchez-Peralta LF, Picón A, Sánchez-Margallo FM, Pagador JB. Unravelling the effect of data augmentation transformations in polyp segmentation. *Int J Comput Assist Radiol Surg*. 2020. <https://doi.org/10.1007/s11548-020-02262-4>.
 - 23. Sánchez-Peralta LF, Picón A, Antequera-Barroso JA, et al. Eigenloss: combined PCA-based loss function for polyp segmentation. *Mathematics*. 2020;8:1316. <https://doi.org/10.3390/math8081316>.
 - 24. Vázquez D, Bernal J, Sánchez FJ, et al. A benchmark for endoluminal scene segmentation of colonoscopy images. *J Healthc Eng*. 2017. <https://doi.org/10.1155/2017/4037190>
 - 25. Bernal J, Histace A, Masana M, et al. GTCreator: a flexible annotation tool for image-based datasets. *Int J Comput Assist Radiol Surg*. 2019;14:191–201. <https://doi.org/10.1007/s11548-018-1864-x>.
 - 26. Angermann Q, Bernal J, Sánchez-Montes C, et al. Towards real-time polyp detection in colonoscopy videos: adapting still frame-based methodologies for video sequences analysis. In: Computer assisted and robotic endoscopy and clinical image-based procedures. 2017. p. 29–41.
 - 27. Silva JS, Histace A, Romain O, et al. Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. *Int J Comput Assist Radiol Surg*. 2014;9:283–93. <https://doi.org/10.1007/s11548-013-0926-3>.
 - 28. Bernal J, Tajbakhsh N, Sánchez FJ, et al. Comparative validation of polyp detection methods in video colonoscopy: results from the MICCAI 2015 endoscopic vision challenge. *IEEE Trans Med Imaging*. 2017;36:1231–49. <https://doi.org/10.1109/TMI.2017.2664042>.
 - 29. Tajbakhsh N, Gurudu SR, Liang J. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Trans Med Imaging*. 2016;35:630–44. <https://doi.org/10.1109/TMI.2015.2487997>.
 - 30. Jha D, Smedsrød PH, Riegler MA, et al. Kvasir-SEG: a segmented polyp dataset. In: Proceedings of the international conference on multimedia modeling (MMM). 2020.
 - 31. Borgli H, Thambawita V, Smedsrød P, et al. Hyper-Kvasir: a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Sci Data*. 2020;7:1. <https://doi.org/10.1038/s41597-020-00622-y>.
 - 32. Sánchez-Peralta LF, Pagador JB, Picón A, et al. PIC-COLO white-light and narrow-band imaging colonoscopic dataset: a performance comparative of models and datasets. *Appl Sci*. 2020;10:8501. <https://doi.org/10.3390/app10238501>.
 - 33. Mesejo P, Pizarro D, Abergel A, et al. Computer-aided classification of gastrointestinal lesions in regular colonoscopy. *IEEE Trans Med Imaging*. 2016;35:2051–63. <https://doi.org/10.1109/TMI.2016.2547947>.
 - 34. Guo Z, Nemoto D, Zhu X, et al. Polyp detection algorithm can detect small polyps: ex vivo reading test compared with endoscopists. *Dig Endosc*. 2020. <https://doi.org/10.1111/den.13670>.
 - 35. Barua I, Vinsard D, Jodal H, et al. Artificial intelligence for polyp detection during colonoscopy: a systematic review and meta-analysis. *Endoscopy*. 2020. <https://doi.org/10.1055/a-1201-7165>.
 - 36. González-Bueno Puyal J, Bhatia KK, Brando P, et al. Endoscopic polyp segmentation using a hybrid 2D/3D CNN. In: Medical image computing and computer assisted intervention – MICCAI 2020. MICCAI 2020. Lecture notes in computer science, vol. 12266. Cham: Springer; 2020. p. 295–305.
 - 37. Qadir HA, Balasingham I, Solhusvik J, et al. Improving automatic polyp detection using CNN by exploiting temporal dependency in colonoscopy video. *IEEE J Biomed Heal Informatics*. 2020;24:180–93. <https://doi.org/10.1109/JBHI.2019.2907434>.
 - 38. Holzwanger EA, Bilal M, Glissen Brown JR, Singh S, Becq A, Ernest-Suarez K, Berzin TM. Benchmarking definitions of false-positive alerts during computer-aided polyp detection in colonoscopy. *Endoscopy* 2021;53(9):937–940. <https://doi.org/10.1055/a-1302-2942>.
 - 39. Misawa M, Kudo S ei, Mori Y, et al. Artificial intelligence-assisted polyp detection for colonoscopy: initial experience. *Gastroenterology*. 2018;154:2027–9. <https://doi.org/10.1053/j.gastro.2018.04.003>.
 - 40. Shin Y, Qadir HA, Aabakken L, et al. Automatic colon polyp detection using region based deep CNN and post learning approaches. *IEEE Access*. 2018;6:40950–62. <https://doi.org/10.1109/ACCESS.2018.2856402>.
 - 41. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition. 2015. p. 3431–40.
 - 42. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. 2018. arXiv. <https://doi.org/10.1109/CVPR.2017.243>.
 - 43. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In:

- Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical image computing and computer-assisted intervention – MICCAI 2015. Lecture notes in computer science, vol. 9351. Springer; 2015. p. 234–41.
44. Chaurasia A, Culurciello E. LinkNet: exploiting encoder representations for efficient semantic segmentation. 2017 IEEE Vis Commun Image Process VCIP 2017 2018-January:1–4. 2018. <https://doi.org/10.1109/VCIP.2017.8305148>.
45. Mahmud T, Paul B, Anowarul S. PolypSegNet: a modified encoder-decoder architecture for automated polyp segmentation from colonoscopy images. Comput Biol Med. 2021;128:104119.
46. Wilson A. Optical diagnosis of small colorectal polyps during colonoscopy: when to resect and discard? Best Pract Res Clin Gastroenterol. 2015;29:639–49. <https://doi.org/10.1016/j.bpg.2015.06.007>.
47. Goyal H, Mann R, Gandhi Z, et al. Scope of artificial intelligence in screening and diagnosis of colorectal cancer. J Clin Med. 2020;9:3313. <https://doi.org/10.3390/jcm9103313>.
48. Patino-Barrientos S, Sierra-Sosa D, Garcia-Zapirain B, et al. Kudo's classification for colon polyps assessment using a deep learning approach. Appl Sci. 2020;10: 501. <https://doi.org/10.3390/app10020501>.
49. Rodriguez-Diaz E, Baffy G, Lo W-K, et al. Real-time artificial intelligence-based histological classification of colorectal polyps with augmented visualization. Gastrointest Endosc. 2020;1–9. <https://doi.org/10.1016/j.gie.2020.09.018>.
50. Jin EH, Lee D, Bae JH, et al. Improved accuracy in optical diagnosis of colorectal polyps using convolutional neural networks with visual explanations. Gastroenterology. 2020;158:2169–2179.e8. <https://doi.org/10.1053/j.gastro.2020.02.036>.
51. Byrne MF, Chapados N, Soudan F, et al. Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. Gut. 2019;68:94–100. <https://doi.org/10.1136/gutjnl-2017-314547>.
52. Berzin TM, Parasa S, Wallace MB, et al. Position statement on priorities for artificial intelligence in GI endoscopy: a report by the ASGE Task Force. Gastrointest Endosc. 2020;92:951–9. <https://doi.org/10.1016/j.gie.2020.06.035>.
53. Ahmad OF, Mori Y, Misawa M, et al. Establishing key research questions for the implementation of artificial intelligence in colonoscopy – a modified Delphi method. Endoscopy. 2020. <https://doi.org/10.1055/a-1306-7590>.
54. Hoerter N, Gross SA, Liang PS. Artificial intelligence and polyp detection. Curr Treat Options Gastroenterol. 2020;18:120–36. <https://doi.org/10.1007/s11938-020-00274-2>.
55. Cheplygina V, de Bruijne M, Pluim JPW. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med Image Anal. 2019;54:280–96. <https://doi.org/10.1016/j.media.2019.03.009>.
56. Golhar M, Bobrow TL, Khoshknab MP, et al. Improving colonoscopy lesion classification using semi-supervised deep learning. IEEE Access. 2021;9:631–40.
57. Pogorelov K, Ostroukhova O, Jeppsson M, et al. Deep learning and hand-crafted feature based approaches for polyp detection in medical videos. In: 2018 IEEE 31st international symposium on computer-based medical systems (CBMS). Karlstad. 2018. p. 381–6.
58. Shin Y, Qadir HA, Balasingham I. Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance. IEEE Access. 2018;6:56007–17. <https://doi.org/10.1109/ACCESS.2018.2872717>.
59. De Almeida TV, Sierra-Franco CA, Raposo AB, et al. Training data enhancements for robust polyp segmentation in colonoscopy images. In: EEE symposium on computer-based medical systems. 2019. p. 192–7.



AIM in Otolaryngology and Head and Neck Surgery

70

Manish M. George and Neil S. Tolley

Contents

Introduction	984
A Brief Introduction into Machine Learning	984
Artificial Intelligence in ENT	986
Limitations, Challenges, and the Future	996
Conclusion	997
References	998

Abstract

There is little doubt artificial intelligence (AI) will define ENT and healthcare as a whole in the near future. Advances in AI in recent years have allowed investigation of a broad selection of ENT datasets including radiological images, sound recordings, neural signals, mechanical measurements, photography, videography, and complex clinicopathological data, among others.

It is important for otolaryngology and head & neck surgeons to appreciate the fundamentals of AI so they can effectively evaluate studies in this field. AI has clear limitations, and yet is subject to marketing hyperbole. A good knowledge base allows otolaryngologists to work through the hype and recognize areas of opportunity in this rapidly growing field. Increasing literacy in data science and machine learning will help with clinical integration of algorithms and promote early adoption of AI tools once safe and effective. It will enable better cooperation with data scientists to direct clinically relevant research. Furthermore, it will facilitate the much-needed multicenter collaboration to take ENT AI research into the future.

This review aims to provide this foundation, illustrated with ENT-specific examples, while highlighting AI's potential benefits, limitations, challenges, and future direction. We discuss a variety of applications in otolaryngology along the main subspecialties from head & neck

M. M. George (✉)
Imperial College NHS Healthcare Trust, London, UK

ENT Department, St. Mary's Hospital, London, UK
e-mail: manish.george@nhs.net

N. S. Tolley
Imperial College London, London, UK

Imperial College NHS Healthcare Trust, London, UK
e-mail: neil.tolley@nhs.net

cancer, thyroid & parathyroid surgery, otology, rhinology, and laryngology.

Keywords

Artificial Intelligence · Machine learning · ENT · Otolaryngology · Head and neck · Otology · Rhinology · Thyroid · Laryngology · Deep neural network

Introduction

Artificial Intelligence (AI) in medicine has seen a paradigm shift in recent years through the merger of technological advancements in powerful graphic processing units, a shift to accessible electronic clinical data and improved multicenter collaboration. Through the automation of tasks previously thought only possible by humans, AI has huge potential to revolutionize healthcare [1]. It is predominantly focused on narrow skills such as prediction, perception, classification, and decision making, with the aim of matching or superseding human level performance.

Within otolaryngology, it is important for clinicians to appreciate the fundamentals of AI so they can effectively evaluate studies in this field.

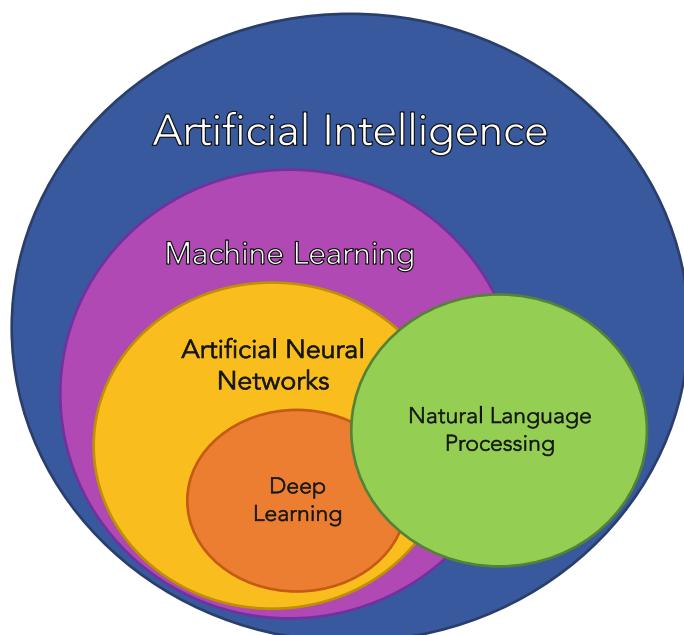
Fig. 1 Artificial intelligence as an area of research includes both NLP and ML. Deep learning is an advanced technique within ML that can also be applied to NLP

AI has clear limitations, and yet is subject to marketing hyperbole. A good knowledge base allows otolaryngologists to work through the hype, recognize areas of opportunity, and contribute positively to research in this rapidly growing field. This review aims to provide this foundation, illustrated with ENT-specific examples while highlighting AI's potential benefits, limitations, challenges, and future direction.

AI encompasses a large area of research, but primarily includes natural language processing (NLP) and machine learning (ML) (Fig. 1). The field of NLP works toward human-level language interpretation and so far is used in personal AI voice assistants and foreign language translation. While the progress in NLP has huge potential in digesting large volumes of unstructured medical text records and interpreting patient speech, ML is the workhorse of current medical AI research.

A Brief Introduction into Machine Learning

Machine learning refers to a group of data orientated, computing and mathematical techniques that enable predictions on new information based on previous experience or data. It enables



learning without explicit programming. The complexity of these algorithms, their ability to self-optimize alongside absence of programmed human knowledge allows the detection of patterns often imperceptible to humans [2]. This makes machine learning one of the most exciting areas of development within medical research.

The basic template for ML involves obtaining a suitably large dataset, preprocessing this data, and dividing it into training and testing parts (Fig. 2). The training data is inputted into the model and through multiple iterations the algorithms' parameters are optimized. This "trained" algorithm is then validated against the unseen testing dataset. While not always undertaken, external validation on data independent to the research center(s) is highly desirable. The three key paradigms within machine learning are supervised machine learning, unsupervised machine learning, and reinforcement learning:

Supervised machine learning involves the development of a predictive model using labeled data, that is, data with known outcomes. The paired input and outputs allow the learning algorithm to optimize by minimizing the mismatch in prediction between inputs and its label, done through many iterations. This algorithm or "function" can then make inferences based on previously unseen data. Supervised learning in medical-based AI is predominantly focused on classification or on regression. A common example is in the classification of images, for example,

cat versus dog, or in medical diagnosis, cancer versus benign [3]. A key component of supervised learning is the "ground truth" used for these labels. The more objective a measure of "truth," such as final histology, the more opportunity for an algorithm to accurately map the real world and make better predictions.

Unsupervised machine learning involves unlabeled data. The algorithm self-organizes without a predefined outcome to discover commonalities and subsequently define "clusters" within data. This clustering may be used to identify previously undefined subgroups of patients in large multivariable datasets. These populations may demonstrate similar prognosis or responses to certain treatments, which in the future may benefit more specific or directed management.

Reinforcement learning is less commonly employed in medical AI research, but after a series of actions, uses a "reward/punishment" type approach depending on final outcome. It is limited due to the real-time nature of the algorithm improvement process, high data volume requirements, and usually slower results. It is frequently used in autonomous game-based AI and was the principle algorithm for AlphaZero, the famed and successful boardgame program. Reinforcement learning has been used in a hybridized approach, alongside supervised learning for developing "dynamic treatment regimes" in graft versus host disease [4]. It has promise in surgical robotics and in clinical decision support-based algorithms,

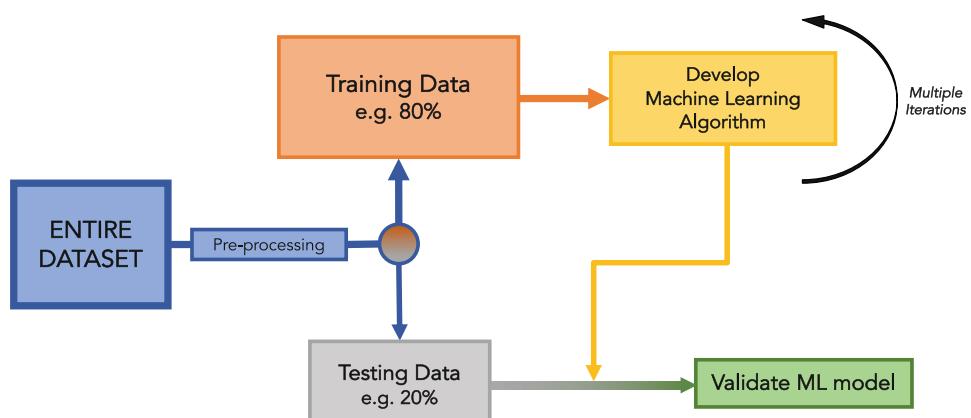


Fig. 2 A simplified overview of the development of an effective machine learning algorithm

especially in critical care [5]. Due to the innate exploratory nature of reinforcement learning as seen with AlphaZero, it has huge potential in uncovering truths and methods in medicine that are not limited by human level cognition nor knowledge.

Perhaps the most promising technique within ML is deep learning (DL) through “deep neural networks” (Fig. 1). A prerequisite to deep learning is the construction of artificial neural networks (ANN) which are inspired by the neural networks of the brain (Fig. 3) [6]. These are highly adaptable systems with good application in interpreting complex heterogeneous data in a highly precise manner. These systems comprise of input and output layers composed of nodes or neurons with bridging “hidden” layers. The complexity of an ANN is in general dependent on the number of hidden layers and a “deep neural network” (DNN) refers to an artificial neural network with multiple layers. Information travels through these nodes with each connection representing a synapse and are attributed a random “weight” or “bias.” These weights or biases are progressively adjusted through multiple iterations to improve the performance of the algorithm. In supervised neural networks, this is achieved through

involved mathematical techniques such as “back-propagation” to minimize error by comparing actual and expected outputs. Across many nodes and millions of connections these algorithms become unfathomably complex but can achieve human or super-human processing capabilities. An important tool in DL is the convolutional neural network (CNN), which are frequently used as an effective device in medical image processing.

Machine learning and specifically deep learning-based AI has demonstrated its utility in medical imaging. Key highlights in this area include: mammogram interpretation in breast cancer [7], skin cancer detection [8], and retinopathy assessment and grading [9] – all at levels comparable or better than medical experts.

Artificial Intelligence in ENT

In otolaryngology, a vast and broad landscape of AI-based research has taken place with number of publications exponentially increasing in recent years (Fig. 4). Machine learning has been employed to detect, classify, diagnose, predict, categorize, and enhance. These successes have

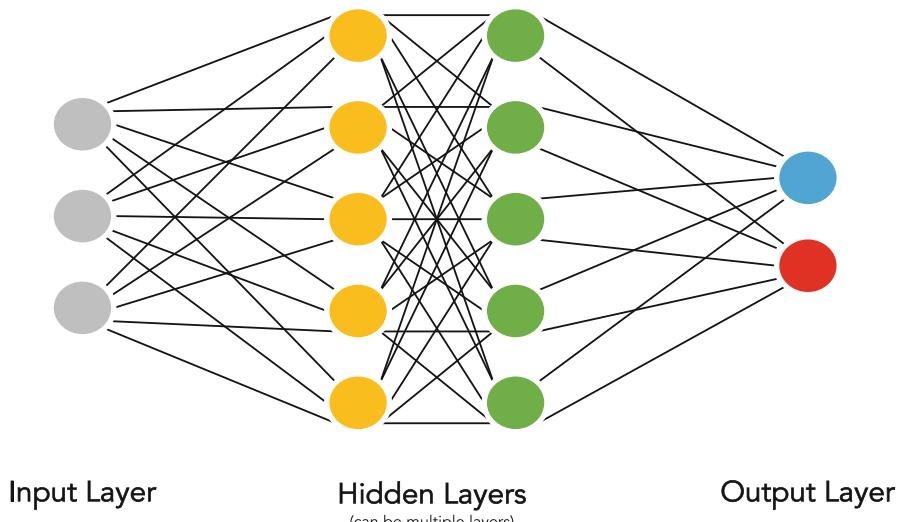


Fig. 3 Modeled on the brain, data passes from the input layer (preprocessed training data) through (multiple) intermediate or hidden layers through connections. Information eventually passes to an output layer and by comparing

actual and expected outputs the weights and biases of these connections are adjusted to improve predictive capacity

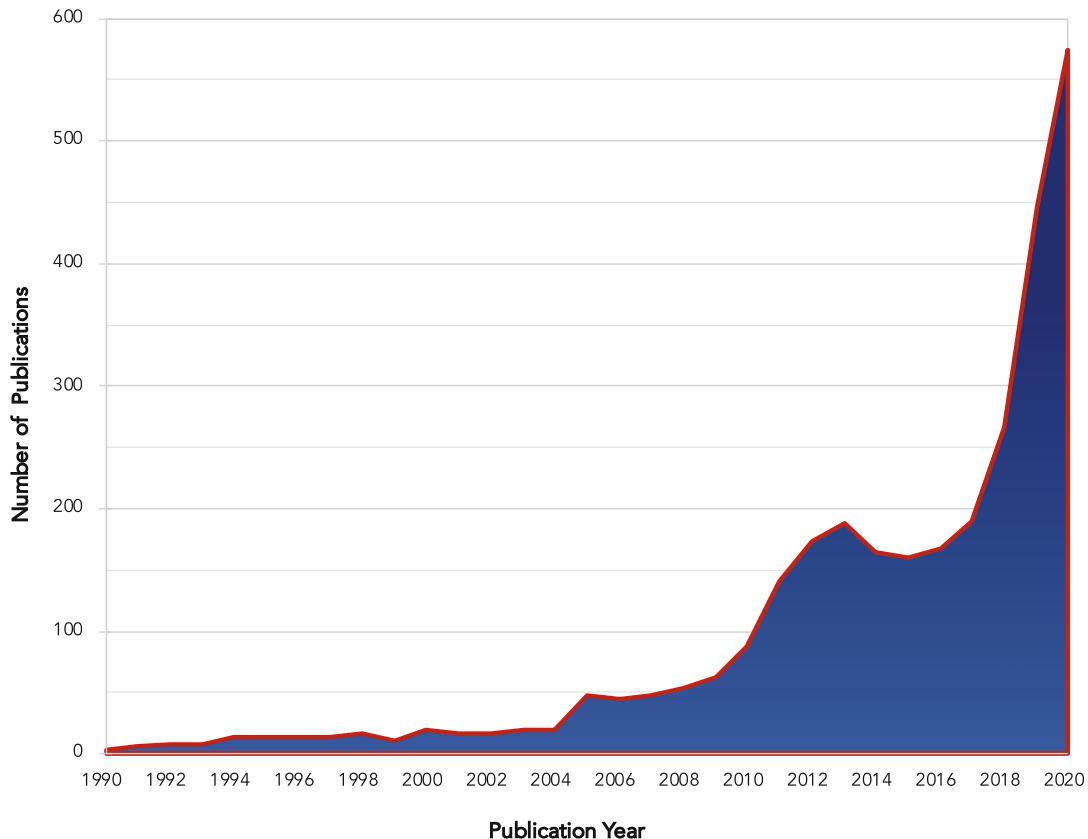


Fig. 4 Publications by year within Medline related to [ENT, Otolaryngology, Head & Neck Cancer, Thyroid, Otology, Rhinology OR Laryngology] AND [Artificial Intelligence, Machine Learning OR Neural Network]

been undertaken in a huge variety of single and hybrid datasets from imaging, genomics, clinicopathological data, sound recordings, neural signals, and mechanical measurements. This review will divide discussion along the main subspecialities within ENT: Head & Neck cancer, Thyroid & Endocrine, Otology, Rhinology, and Laryngology.

Head and Neck Cancer

AI research in Head & Neck cancer is presently the principal area of investigation within Otolaryngology. Large amounts of accessible data in imaging, pathology, genetics, and proteomics lend itself to an ML-based research approach.

Radiomics

Radiomics refers to the extraction of a large number of features from radiographic medical images and has now developed into a leading area of AI

investigation. Multidimensional medical images contain far more information than is detectable by humans. These imperceptible patterns may represent “image biomarkers” which may correlate with characteristics of the disease process, predicting both diagnosis and prognosis.

Predicting Outcomes: The largest area of AI research in head and neck cancer is in radiomics-based prognostication including survival, locoregional recurrence, and metastasis prediction [10]. With wildly varying survival rates in this heterogeneous population of tumors, better predictive tools are highly desirable. In the last 4 years, over 50 studies have been conducted in this area, investigating differing imaging modalities and distinct subsets of head and neck cancer. Shen et al. found a hybridization of MRI-based radiomic models with staging and EBV tumor status best predicted progression free survival in nasopharyngeal cancers [11]. Liu et al. demonstrated a superior

predictive power of overall and disease free survival when combining pre- and posttreatment PET-CT radiomic modeling with clinicopathological features in head and neck squamous cell cancer [12]. This predictive success however has not always been demonstrated in published MLA radiomic-based studies [13]. With more accurate prognostication, better-informed patients and clinicians may opt for more or less aggressive treatment in a personalized management approach.

Predicting Treatment Response: Related to prognostication is the assessment of treatment response. Response to chemotherapy in sinonasal and nasopharyngeal cancers has been predicted using algorithms trained on pretreatment MRIs [14, 15]. Zhai et al. using a neural network trained on a combined dataset of clinical and radiomic features, was one of the first groups to predict individual lymph node failure pretreatment [16]. Models such as these may allow better pretreatment planning and subsequent intensified or directed therapy, which may include neck dissection or higher dose radiotherapy.

Prediction of Pathological State: Automated staging and prediction of tumor histological features have also been researched extensively in HNSCC. Multiple studies have been able to predict histological features, grade, and degree of differentiation in both CT [17, 18] and MRI [19].

Overall, developing head and neck cancer radiomics may result in superior prediction in survival, treatment response, and earlier pathology diagnosis. This may in turn result in more effective individualized management for these cancer patients.

Radiological Staging

MLAs have been able to accurately stage advanced laryngeal tumors at T3 versus T4 through a contrast enhanced CT image data set [20]. Romeo et al. in a smaller subset of CT images for oral cavity SCC, through machine learning classifiers was able to demonstrate autonomous staging of both T and nodal status [21]. Tomita et al. recently trained a deep learning neural network to identify cervical nodal metastasis on CT in oral SCC [22]. They showed an AUC significantly better than that of the radiologist

comparator. A similar, but older, study in preoperative HNSCC was able to diagnose pathological nodes and was the first to detect extranodal extension on external validation with a high AUC of 0.91 [23]. Both of these studies highlight the important role deep learning algorithms may have in radiology support for preoperative diagnosis, risk stratification, and directed management. In the near future AI tools with high sensitivity can flag certain high-risk scans for earlier review by an experienced radiologist and thus streamline diagnosis.

Clinical Head and Neck Oncology

Predicting Treatment Toxicity: Post radiotherapy toxicity of the head and neck can manifest in a multitude of ENT symptoms: dysphagia, hearing loss, osteoradionecrosis, and xerostomia. Risk of xerostomia in particular has been examined in multiple studies; Jiang et al. used supervised machine learning methods on a combined dataset of sociodemographic, clinicopathological features, and radiation dose distribution [24]. They were able to accurately predict risk of xerostomia and also identify which salivary gland regions correlated most with this complication.

Auto-segmentation: In head and neck clinical oncology, deep learning networks have been used to automate segmentation of specific anatomical structures of interest to facilitate radiotherapy planning [25].

Nearly all studies examining the value of machine learning models in radiomics for head and neck cancer focus on a single modality. It is possible if not probable that combination of multiple imaging modalities may further improve accuracy of prediction. In late presenting and difficult to access tumors such as skull base and sinonasal, a highly accurate radiomic-based diagnostic and prognosticating algorithm would be invaluable in managing these patients.

Histopathology

Much like in radiology, histopathological images can be interpreted using CNNs. Halicek et al. digitized whole slide pathological images of head and neck SCC and interpreted a test set with a high AUC of 0.92 [26]. Each slide was

broken down into multiple small “image patches” to enable reliable training and subsequent interpretation. Results however were limited by image artifact including out of focus regions on the whole slide photograph. Additionally, due to “image patch” size, minimum interpretable SCC size was no larger than approximately 10 cells.

Ultimately, the use of computer diagnosis in histopathology, through deep learning methods, may rapidly speed up diagnosis. In the near future it may augment the histopathologist rather than replace them by flagging “high-risk” specimens for earlier formal review. Later, once more reliable and robust algorithms are produced on large volumes of multicenter data it may serve as diagnostic fail safe.

Multispectral Imaging

Multispectral or hyperspectral imaging has emerged as a tool to enhance issue identification via assessment across a broad range of both invisible and visible light. Combining the data images from different wavelengths multiplies the amount of information available and applying artificial intelligence to this dataset may uncover associations not obvious to the human eye or brain.

In a relatively small sample size, a team from Stamford University through a supervised learning “naive Bayesian” classifier demonstrated reasonably accurate identification of oropharyngeal malignancies on multispectral images taken from flexible nasopharyngoscopy [27]. Use of deep learning algorithms to identify normal tissue type on fresh frozen human cadaveric specimens has been undertaken with success. A model developed from multispectral imaging on operating microscope photography was superior to otolaryngology residents in diagnostic accuracy (82% vs 70%) [28]. This algorithm was also significantly better than one built on white light alone and in the future may enhance real-time intraoperative tissue assessment. In another study, head and neck surgical specimens containing marginal tumor tissue had hyperspectral images taken within 2 h of excision. These were analyzed with a machine classification model. The final algorithm successfully distinguished cancer from normal tissue in a variety of

head and neck malignancies at an accuracy of 90% [29].

ENT widely uses image representation devices including rigid and flexible endoscopes, microscopes, and robots. These algorithms may improve automated detection in malignancy, and further development may allow real-time assessment of tumor margins during surgery. Ultimately these systems may enhance intraoperative surgical vision in other ENT subspecialties through better visual definition of a variety of tissue types.

Genetics and Molecular Markers

Large genetic and molecular data sets are suitable for MLA application. Complex and as yet unknown or unknowable associations between genes and or proteins may code for tumor biology and behavior.

Gene Expression: Some fascinating AI genetic prediction models have been developed in head and neck cancer. Stepp et al. developed a 40 gene model from the genomics of the primary tumors of HPV+ve oropharyngeal SCC [30]. This algorithm, based on the gene expression of these tumors alone, was able to successfully predict nodal metastasis with an AUC of 0.93. An earlier study looked at combined datasets of clinicopathology and genomic data. They applied feature selection and machine learning to develop a model that improved prognosis prediction in oral cancer beyond clinicopathological features alone [31].

Molecular Markers: Carnielli et al. used protein and peptide biological signatures to predict lymph node metastasis in oral SCC [32]. They did this with an accuracy above 74%, validated on a test set. Using a support vector machine learning algorithm Bohnenberger et al. were able to identify “proteomic diagnostic biomarkers” to differentiate between primary lung SCC and HNSCC metastases. This model, based on the analysis of over 1000 proteins, demonstrated an accuracy of approximately 87% on an independent validation set [33].

Unique protein and gene expression combinations may represent definite subgroups of cancer patients. These specific patterns may identify patients with more aggressively behaving tumors or those with lymph node metastasis irrespective

of imaging. A more accurate stratification would guide more or less aggressive treatments.

Thyroid and Endocrine Surgery

Thyroid Cancer

In a diagnostic sense, thyroid cancer differs from most other head and neck cancers primarily due to the frequent indeterminacy of the principal investigations. A large number of nodules, approximately 15%, are classified as indeterminate despite further invasive testing including ultrasound and fine needle aspiration cytology. The recommendation in these cases is for diagnostic hemithyroidectomy to obtain final histological diagnosis. Reliance on surgery for diagnosis, which carries morbidity, is a key area for improvement. Improved accuracy for presurgical investigations can streamline care for patients and avoid unnecessary surgery. Furthermore, the increase in detection and subsequently incidence of thyroid requires better risk stratification and prognostication to guide individualized treatments.

Thyroid Ultrasound: AI-based interpretation of thyroid US images can be used to reduce the recognized subjectivity in image assessment. Buda et al. produced a deep learning model that demonstrated comparable accuracy for diagnosis on ultrasound thyroid images, when compared to 9 radiologists with an AUC of 0.87 and 0.82, respectively [34]. A more recent study used a “similarity based” MLA to accurately classify US nodule images as benign or malignant. This algorithm used its labeled training database to identify the “closest” appearing image to the new test image [35]. As the training images were labeled on pathological ground truth, this algorithm was able to provide partial “explainability” to its output, thus minimizing a key concern of “black box” AI neural networks (discussed later).

Importantly, there may be as yet unknown subtle patterns in US images that correlate with malignancy. With large enough datasets, it may be possible to deliver AI-based interpretation that diagnoses so-called indeterminate nodules as benign or malignant with superior accuracy. This may revolutionize the management of such

thyroid lesions by reducing the need for diagnostic surgery.

MRI Radiomics: Beyond US, MRI has been used to predict the aggressiveness of papillary thyroid carcinoma (PTC) preoperatively. Wang et al. imaged 96 patients prior to surgery for papillary thyroid carcinoma and using ground truth histopathological features, trained an MLA to differentiate between aggressive and non-aggressive PTC [36]. The algorithm based on the radiomic features alone demonstrated an impressive AUC of 0.92 versus an AUC of 0.52 for prediction based on clinical and patient factors alone. Wei et al. used a similarly trained MLA to predict extra-thyroidal extension on preoperative MRI scans [37].

MRI is not a standard preoperative assessment tool in thyroid cancer; however, with powerful predictive AI tools extracting reliable diagnostic patterns from this data, better preoperative planning and more individualized treatment plans for PTC may be possible.

Clinicopathological Prediction: Historically, prognostication in thyroid cancer has been based on TNM, AMES, and MACIS classifications, all devised on multivariate statistical analysis. MLAs however have proven a powerful tool in predicting survival in thyroid cancer patients. Using only seven clinicopathological variables, based on the very large national US “Surveillance, Epidemiology, and End Results” database, Mourad et al. predicted death within 5 years with an accuracy of 96% [38].

Cytopathology Diagnosis: Fine needle aspiration (FNA) cytopathological interpretation is a complex task, in part because of the large slide images with relatively isolated or sparse areas of interest. Information is too diffuse for reliable diagnosis based on whole slide images with a traditional single classification algorithm. For this reason, double MLAs have been applied to cytopathological slides, the first to identify groups of follicular cells (screening algorithm) the second to classify these groups (classifying algorithm) [39]. In a dataset of 908 FNA slide sampled the MLAs predicted malignancy with an AUC of 0.932, comparable to the cytopathologist AUC of 0.931. The near instant interpretation of

cytopathology may work to highlight “high-risk” slides earlier to the pathologist who can better direct efforts to return earlier diagnosis in patient with potential malignancy.

AI developments in thyroid cancer may not only work as an adjunct to human experts in streamlining diagnosis but later could help revolutionize the MDT through more accurate diagnosis and better risk stratification.

Parathyroid Surgery

Intraoperative Assistance: In primary hyperparathyroidism, the intraoperative evaluation of pathological glands can be difficult. A relatively new technology of near infra-red imaging to assess auto-fluorescence of parathyroid glands is a promising intraoperative adjunct. A machine learning multiclass decision tree using auto-fluorescence intensity, gland volume, and “heterogeneity index” was able to differentiate between abnormal and normal glands with an AUC of 0.98 [40]. This algorithm was not prospectively validated and was developed on a relatively small pool of 106 patients; however, it demonstrates potential in objective intraoperative assessment without need to excise the gland. Larger studies with robust validation are desirable.

A study from Germany in 2020 used intraoperative hyperspectral photography to differentiate between thyroid, parathyroid, and recurrent laryngeal nerves within less than 1.4 s of processing. A support vector machine learning algorithm based on annotated pictures was able to identify parathyroid tissue with sensitivity of 65% and specificity of 94% [41]. The dataset involved only nine patients and larger studies are warranted. In the future, more sensitive and more extensively validated algorithms, with similar efficiency, may revolutionize intraoperative surgical vision in thyroid and parathyroid surgery.

Preoperative Diagnosis: Using demographic, clinical, and laboratory measures Imbus et al. developed a random tree model of MLA to predict multi-gland disease versus single adenoma in primary hyperparathyroidism at an accuracy of 94% [42]. Interestingly, they adjusted the algorithm to maximize positive predictive value for multi-gland disease at 100%. Using prediction software

such as this we may identify where 4 gland exploration is almost certainly necessary, thus streamlining tertiary center referral and possibly avoiding the time, cost, and radiation associated with possible revision surgery and investigations such as sestamibi scanning.

Otology

You et al. published an excellent and thorough review of AI in otology in 2020 evaluating 38 studies [43]. The scope of this review is to give an overview of ENT as a whole but will touch upon some of the key otology areas. There are five broad divisions that AI research has focused on within otology: Imaging diagnostics and radiomics, auditory brainstem response analysis, sensorineural hearing loss prediction, hearing impairment technologies, and balance pathology. We provide a summary of AI research in these areas. As with other fields, a large amount of otology AI research has been undertaken in imaging modalities. A unique and extensively investigated area has been in hearing assessment and rehabilitation, driven in part by the hearing industry’s commercial interest in technological advancement.

Imaging Modalities and Radiomics

As in other fields, AI-based analysis of imaging has been applied in otology to otoscopic images, CT scans, and even optical coherence tomography.

Radiology: Accurate automated segmentation and identification of key temporal bone structures on CT scans will streamline and augment preoperative planning for otologist and lateral skull base surgeons. Fauser et al. used deep learning-based methods, on ground truth of expertly annotated images, to segment temporal bone CT scans [44]. Tested on images from 24 patients, they highlight that using this accurate, fast, and fully automated approach may facilitate planning in procedures such as cochlear implantation and vestibular schwannoma surgery.

Wang et al. developed dual DL algorithms to localize the region of interest on temporal bone CT scans and then further classify the images into chronic otitis media (COM) or normal at an AUC equivalent or superior to experts [45]. The classification algorithm also demonstrated superior

recall rates in differentiating chronic suppurative otitis media and cholesteatoma at an accuracy of 77% versus 74% in clinicians. Not only was this model more accurate but also had better diagnostic consistency when compared to experts.

Novel Modalities: The detection of endolymphatic hydrops in Meniere's disease via cochlear optical coherence tomography (OCT) has significant potential as a noninvasive and useful investigation [46]. In a mouse study and using histology as ground truth, Liu et al. trained a CNN to identify both normal and hydrops-related Reissner's membranes on OCT images [47]. On a small validation set of 37 cochlear images they demonstrated a high accuracy of 92%.

Radiomics: Within ENT, radiomic-based prediction and diagnosis is not restricted to the domains of head and neck cancer. Sensorineural hearing loss (SNHL) secondary to chemoradiotherapy for head and neck cancer was predicted through application of machine learning algorithms to CT images of the cochlea. Of the 490 identified radiomic features, 10 were associated with SNHL and were able to predict SNHL in this patient group at an impressive accuracy of 70% [48].

Oto-endoscopic Images: Misdiagnosis in medicine is not uncommon and is seen frequently on otoscopy in the community. It can result in delayed and unnecessary referrals leading to patient anxiety, avoidable cost, and even morbidity. Furthermore, even among experienced clinicians, otoscopic examination is prone to interobserver variability [49]. Objective computer vision tools may minimize this subjectivity and multiple studies have examined the use of MLAs in interpreting oto-endoscopic images. Lee et al. using a dataset of 1338 tympanic membrane images developed a CNN which was able to recognize the laterality of eardrum with 99% accuracy and identify perforation with an accuracy of 91% [50]. Similarly, Wu et al. developed a convoluted neural network on a larger dataset of 10,000 pediatric oto-endoscopic images [51]. The model was able to diagnose and differentiate between acute otitis media, otitis media with effusion, and normal tympanic membranes. While not externally validated, the model had a very high accuracy on an internal test dataset of 97%.

The development of highly accurate and broadly trained algorithms may revolutionize diagnosis in the community, especially in untrained individuals, thus speeding and streamlining the referral pathway. This has significant implications in the delivery of effective diagnostic care in the developing world where expertise in otolaryngology maybe at a premium.

Auditory Brainstem Response Interpretation

Auditory brainstem response (ABR) is considered an objective assessment for evaluating hearing thresholds as a voluntary response is not required. Nevertheless, the interpretation of the detected traces is subjective and relative to clinicians' expertise and judgment. Studies dating back at least 15 years have used artificial intelligence to classify ABRs more objectively. Most recently in 2019, a DNN was trained to classify paired auditory brainstem response waveforms into three categories, specifically "clear response," "inconclusive," or "response absent" with an accuracy of 93% [52]. Like with other complex datasets, deep learning can retrieve specific patterns within the ABR reading that may be unknown or poorly defined by clinicians. Ultimately reliable deep learning-based algorithms may diagnose in real-time and streamline ABR or other electronic signal-based screening programs.

Sensorineural Hearing Loss Prediction

Predicting audiology outcomes in patients with hearing loss is difficult given heterogeneity of disease and the presumed complex interplay between large variety of clinical, demographic, and pathological factors. This is especially the case in sudden sensorineural hearing loss (SSNHL). In 2018 a relatively large cohort of 1220 patients with SSNHL, Bing et al. investigated several MLAs to predict hearing recovery [53]. Using 149 separate clinicopathological variables they demonstrated that a deep learning-based algorithm (deep belief network) was the best predictor of audiological outcome at 78% accuracy. More recently in China, MLAs were employed to predict the likelihood of noise-induced hearing loss (NIHL) in a collection of

factory workers using several clinico-demographic factors alongside full-shift noise measurements. They were able to quantitatively predict NIHL at an accuracy of 80% [54]. Models such as this may better risk stratify a workforce, highlight susceptible individuals, and thus protect employees from occupational NIHL.

Hearing Impairment Technologies

Hearing Aids: A multitude of factors including demographic, cognitive, and audiotrical contribute to speech intelligibility. Kates et al. proposed a neural network plan to predict intelligibility of speech in hearing aid users based on clinico-demographic factors with an intention to enhance the audibility of high-frequency speech [55]. A major hurdle within hearing aid development is minimizing speech degradation. Enhancing speech sound while mitigating background noise has been successfully achieved through use of ANN. “Perceptual evaluation of speech quality” improved by 24% when a deep learning system was applied to recordings of speech sounds in a noisy office environment [56].

It is not unreasonable to expect MLAs to be developed in the near future that allow clean and enhanced audio streamed directly to device-paired hearing aids. Similarly, once processed speech is enhanced to high level intelligibility, a next possible step is near real-time and faithful translation of foreign language audio to headphones or hearing aids.

Cochlear Implants: AI has been introduced to a variety of areas within cochlear implantation and usage. Deep learning in particular will be paramount in the development of a customized cochlear implant programming.

Cochlear implant (CI) users describe limitations in the appreciation of music, especially in the perception of vocals. As recently as 2020, Tahmasebi et al. developed a source separation algorithm using a deep neural network to enhance music enjoyment in individuals with CIs [57]. This was done through effectively enhancing voice sound relative to background instruments. Most impressively, the study allowed real-time assessment of this deliberate “music remixing” in real-world environments with ambient sound.

Evoked compound action potentials (ECAPs) are useful in both postoperative and intraoperative assessment of CI and cochlear nerve function. Analysis of electrically evoked compound action potentials has been investigated using neural networks across several studies. Gartner et al. evaluated the previously developed “Automatic Neural Response Telemetry” (AutoNRT™) software, which demonstrated comparable accuracy to audiologists with a relatively fast assessment time of under 1 min per electrode contact [58]. The ECAP is clouded by interference from the implanted device itself. Extracting the neural response component of the trace has been done through support vector machines which have been shown capable of differentiating normal hearing signal and artifact [59].

An interesting study by Pile et al. developed a simulated robotic system that detected cochlear implant electrode “tip fold-over.” Through a variety of force feedback readings on a synthetic cochlear model, a support vector MLA was able to accurately predict this known intraoperative complication in real-time at 88% accuracy [60]. Rather than relying on postoperative assessment or “on table” fluoroscopy, this tool can permit immediate intraoperative rectification.

Transcription Apps: Engineered by AI natural language processing models, speech to text transcription apps have already proven their worth for patients reliant on lipreading due to the use of COVID-19 related face masks.

Balance and Vestibular Pathologies

A decision support system for balance disorders was created through use of decision tree-based algorithms trained on clinical data from just under 1000 patients with up to 350 variables. It has a reported diagnostic accuracy rate of 59.3 to 89.8% for general practitioners (GPs) and from 74.3 to 92.1% for balance experts [61]. Krafczyk et al. developed an ANN to diagnose underlying balance disorders based on posture or “body sway” [62]. They were able to identify normal balance, postural phobic vertigo, anterior lobe cerebellar atrophy, primary orthostatic tremor, and acute unilateral vestibular neuritis with high sensitivity and specificity using only

posturography measurements. In using both clinico-demographic datasets and worn device measurements AI can improve diagnostic accuracy in balance disorders for both GPs and balance specialists.

Rhinology

AI in rhinology has not been investigated to the same degree as head & neck oncology and otology. This is in part related to the sheer volume of accessible data in cancer imaging, genetics, and audiology. However, along with the common imaging and pathological interpretation, endotyping of chronic rhinosinusitis has been investigated considerably.

Imaging Diagnosis

Imaging, as mentioned previously, lends itself well to CNN analysis. Two separate studies have successfully used a CNN, Google Inception-V3, to interpret CT sinus scans for osteomeatal complex occlusion [63] and anterior ethmoidal artery identification [64]. Chowdhury et al. examined CT scans in 239 patients with chronic rhinosinusitis (CRS). Augmented to 956 images, they demonstrated algorithm accuracy of 85% and AUC of 0.87 in defining the osteomeatal complex as open or closed. Huang et al. in a larger data set of 388 patients was able accurately highlight the anterior ethmoidal artery on images with an accuracy of 82.7% and AUC of 0.86. Another study, using only CT coronal slices, achieved an 81% accuracy in their neural network model for diagnosis of concha bullosa [65].

Due to a degree of subjectivity and interobserver variability associated with interpretation of CT sinus imaging, developing objective, rapid, and highly accurate assessment of these images is desirable [66]. CNNs may eventually provide this as they demonstrate promise in interpreting and highlighting of key components on CT sinuses.

Pathological Diagnosis

A far less studied area of image-based deep learning is in digitized images of rhino-cytology and histology preparations. CNNs have been trained to differentiate between eosinophilic and non-eosinophilic nasal polypsis on histology

[67]. Their training dataset of 167 patients' slides was augmented to a robust 23,048 separate images or "patches" from slide image breakdown and externally validated with high diagnostic accuracy. Streamlining the time-intensive nasal cytological diagnostics has also been attempted with a CNN composed of three block layers, showing promising results [68]. Here, Dimauro et al. produced a very reliable model for cell identification and cell classification which was between 98 and 100% accurate on validation depending on the cell type. These developments may eventually lead to useful and rapid outpatient diagnostics.

Chronic Rhinosinusitis Endotyping

Outside of neural networks, AI has been used in assessment and clinical categorization for patients with chronic rhinosinusitis (CRS). CRS is associated with a large worldwide disease burden affecting over 100 million people [69]. Furthermore, CRS consists of multiple differing poorly understood biological subtypes or endotypes that may correspond to distinct pathophysiologies. These different groups may respond with varying efficacy to different treatments. Defining CRS as a two-group phenotype, that is, with or without nasal polyposis may be an oversimplification and suboptimal for effective individualized care. Several papers have undertaken unsupervised clustering analysis of CRS to predict endotype groups and subsequently predict surgical benefit, treatment response, and need for revision surgery.

Soler et al. used clinical data (demographics, Sino-Nasal Outcome Test-22 (SNOT-22), CT scoring, olfactory testing, and endoscopic scoring) from 690 patients who had initial failed medical management CRS. They were able to define five clusters, three of which predicted better surgical outcomes on SNOT-22. The other two clusters demonstrated no benefit to surgery over continued medical management [70].

Liao et al. used principle component analysis to define seven distinct clinicopathologic multidimensional clusters of CRS [71]. They assessed a total of 67 total variables, 28 clinical ranging demographics, symptomatology, endoscopic findings, and CT scoring alongside 39 intraoperative

molecular and cellular markers. In this way they were able to characterize difficult to treat CRS and predict differing response to treatment.

Laryngology

Assessment of the larynx, voice, and swallow has had limited dedicated AI research. Nevertheless, many of the studies undertaken have been unique, interesting, and offer significant promise.

Voice and Larynx

Most of the promising AI-based laryngology-related research has focused on diagnosis using noninvasive digitized media-based analysis. Much like image data, the large amounts of information held within video or audio recordings lend itself to interpretation by MLAs.

Anecdotally, expert voice clinicians may demonstrate an intuition in classifying voice pathology prior to endoscopic examination. Presumably, there are pathognomonic features detected by experienced humans that may not be easily expressed in words. It is also possible that subtleties within the audio recording data may code complex patterns imperceptible to the human ear that will allow highly accurate superhuman noninvasive diagnosis.

Audio-Based Prediction: Using ANNs to analyze voice recordings has been conceptualized as far back as the year 2000 [72]. In 2018, a remarkable study by Fang et al. demonstrated superior performance of deep neural network versus other AI systems in diagnosing “pathological voice” on audio samples. Trained on 462 labeled voice clips the ANN had a 99.2% accuracy when externally validated [73]. In the same year, Cesari et al. showed superior identification of “pathological voice” versus “normal” on voice recordings using a machine learning decision tree algorithm. This was best seen in the subcategories of reflux laryngitis and hyperkinetic dysphonia [74]. This program had been incorporated into a mobile phone app paving the way for highly accessible, noninvasive, AI-based diagnostics in the near future.

Video-Based Prediction: Videographic data can also be interpreted in a similar way. In a relatively small study, high-speed laryngoscopy

video (4000 images per second) has shown significant potential in differentiating T1a tumors from precancerous laryngeal lesions when analyzed by a support vector MLA [75]. The MLA was, in theory, able to identify abnormal mucosal vibration and demonstrate a specificity and sensitivity of 100%, but larger training and testing set would be needed to validate this further. In the future, high accuracy noninvasive tests may limit unnecessary diagnostic surgery in small benign laryngeal lesions.

Clinical Data-Based Prediction: Interestingly, success in voice pathology diagnosis has been demonstrated in using MLAs on simple clinical data without incorporation of digitized media. Tsui et al. applied an ANN to a combined dataset of demographic and symptomatic clinical data to predict underlying pathology for dysphonia [76]. The ANN was the most accurate of the tested MLAs, differentiating between phonotrauma, palsy, and neoplasia at an accuracy of 83%. This is especially impressive given the data set did not include video or audio; however, the model was not externally validated.

Image-Based Diagnosis: In late 2020 “AGATI” (Automated Glottic Action Tracking by Artificial Intelligence) a computer vision tool could identify unilateral vocal cord palsy by tracking true vocal fold movement on fiberoptic laryngoscopy images. Through quantitative assessment of vocal fold movement via surrogate measure of “anterior glottic angles” the program had a high diagnostic accuracy with AUC of 0.87 [77].

Swallow

Swallow evaluation through recognition of imperceptible or abstract patterns in novel noninvasive measurements may revolutionize outpatient or bedside diagnostic approaches.

Dysphagia assessment through mapping hyoid bone displacement on video fluoroscopic swallow (VFS) has been automated with an accuracy of almost 90% using deep learning [78]. This has reduced task time from over 30 min for manual annotation to less than 1 min for the algorithm. Interestingly, the same team subsequently taught an MLA to accurately predict hyoid bone movement using vibration signals from a worn neck

accelerometer, using human annotated VFS as ground truth [79]. Using similar principles, identification of upper esophageal sphincter opening via vibratory output was also possible with over 90% accuracy [80, 81]. This algorithm-based vibration assessment of swallowing has large potential as adjunctive noninvasive and zero radiation diagnostics for dysphagia.

Limitations, Challenges, and the Future

Data Access, Data Processing, and Further Research: While the charm of machine learning algorithms lies in their ability to manage large volume complex heterogeneous datasets, their successful development is dependent upon data that is clean, often annotated, appropriately specific, or broad and sizeable enough. This volume of data can both be difficult to obtain and time onerous to pre-process. Furthermore, while collaboration is improving, the data sources for the vast majority of studies are single institution – even in well-studied areas such as head and neck oncology.

Otolaryngology needs cooperative large-scale multicenter data for training and additional comprehensive external validation. Additionally, the majority of machine learning research uses single modality data inputs. Multimodal imaging-based AI is in its early stages, but has potential to further improve diagnostic and prediction capacity. Beyond using imaging alone and mimicking healthcare professionals' approach, the merger of demographics, clinicopathology, biochemistry, pathology, and imaging among other data may lead to holistic and powerful AI assessment tools.

While not obvious, ENT surgeons interested in developing real-world application tools need to incorporate unequal error costs to their models. This is central to assessment in cancer. Rather than just differentiating between malignant and benign the impact of a false negative must be recognized and factored into any algorithm. Subtly minimizing false negatives at the expense of false positive rates is instinctively what clinicians do and any ML model should be adjusted as such.

Otolaryngologists interested in artificial intelligence should familiarize themselves with

international guidance on both the interpretation of AI research and design of AI studies [82–84].

Explainability: Through the fine tuning of millions of connections (especially in deep learning), these algorithms become unfathomably complex. Machine learning algorithms are often coined as “black box” models, where data is introduced and an output produced, without an understanding of the decision process. Irrespective of their superhuman evaluation abilities, the lack of explainability and subsequent accountability when employing these tools make humans uncomfortable. Medical professionals in particular need to trust any device they use in patient care. Lack of explainability will only exaggerate any concerns clinicians have with new AI technologies.

In spite of this ethical dilemma, due to the potential to revolutionize patient care some argue that the “black box” should be overlooked in favor of adopting this technology if demonstrably effective. They highlight that clinicians frequently work on intuition and often are unable to immediately and precisely explain their reasoning nor the evidence base for some clinical decisions they make. Most AI ethicists however promote development of “explainable AI” for legal, technological, patient, and clinician-based reasons [85]. In the distant future, hyper advanced AI may be able to explain complex and multi-dimensional associations in data in a way comprehensible to humans; before this, explainable AI appears a necessary stepping stone for widespread adoption.

Privacy: The production of high-functioning and reliable algorithms is reliant on access to huge volumes of data, normally retrospective. It is impractical to seek consent for thousands or millions of patients and while much image, sound, neural signal related data can be anonymized easily, this is more difficult with clinicopathological variables. In some cases, data that is presumed to be adequately anonymized data can actually be traced back. Robust protocols to protect confidentiality should be adhered to widely.

Fragility: Deep neural networks in particular are prone to “fragility.” Due to the complex nature of these networks they have the ability to extract

subtle features from input data. Some of these patterns are imperceptible to humans and may confer a superhuman level of detection or diagnosis. However, many others may be irrelevant but given salience due to the idiosyncrasies of the training dataset. Subsequently, small and seemingly innocuous changes to the image or data can confuse the algorithm into misclassification.

One method to help overcome this fragility, which will likely be expanded in the near future, is adversarial training. Here, a second algorithm will test and train the original with examples designed to fool it thereby engineering a more robust model to safeguard against mistakes. This method in itself is imperfect and may strengthen algorithms in some areas while weakening it in others. Ultimately however, the vast and deep human experience of the world enables us to see images as a series of concepts in context and not be distracted by abstract or irrelevant minutiae. This is not something narrow AI can achieve easily.

Universalizability: If algorithms are trained on one data acquisition source, for example, a specific ultrasound machine model, they learn based on those particular set of individualized parameters and signals. As a result, they may not be well suited to interpret data from an alternative source, for example, different scanners or different recording devices. In mixed data sources, machine learning can be more difficult. In large enough studies however, researchers can hope an algorithm may see beyond this noise and develop into a more general model with good accuracy irrespective of the instrument used. Conversely, it is recognized that the use of data with widely varying acquisition parameters may cause signal alterations that are not attributable to biology and thus reduce both efficacy of model and limit the algorithm's generalizability.

Patient and Public Involvement: While a thorough understanding of any new technology is not a prerequisite for use, patient confidence in the system is a necessity. Involvement in the public at early and late stages of development will not only help with trust but will also ensure their patient-specific needs are met. Tied in with public trust is the concept of AI supporting doctors rather

than replacing them. In this sense algorithms should be narrow enough to leave the primary responsibility of care delivery with a human clinician, in the near future at least.

Bias: Bias in algorithms remains an area of concern in real-world AI application. Systematic biases within a dataset can become represented within the trained model and result in unfair, unethical, and unreliable outcomes. Due to the lack of explainability in DNNs as outlined above, these biases may be hard to recognize and thus underlines the apprehension around widespread use of such systems.

Outside of medicine, image assessment algorithms in cameras have reflected data limitation and subsequent bias through inability to recognize dark skinned faces and mislabeling of East-Asian faces as blinking. Ethnic minorities are usually underrepresented in large research studies in the developed world and outside of AI this contributes to poorer risk stratification, diagnostics, and management. Skin cancer detection algorithms prove significantly poorer at identifying malignancies in darker skin and is likely a product of the data source containing predominantly light skin images [86]. Whatever the reasons, developing and employing algorithms that have lower rates of efficacy along lines of ethnicity or any other group raises serious ethical questions.

In other ways, biases within data can result in unexpected algorithmic quirks. In two large image diagnostic studies, the likelihood of diagnosis skin cancer or pneumothorax was increased with the presence of a ruler [87] or chest drain [88], respectively. Whilst perhaps less sinister in its social equality implications, these flaws still undermine accuracy and faith in complex AI systems.

Conclusion

Advances in artificial intelligence in recent years has allowed complex examination of a broad selection of ENT datasets including radiological images, optical coherence tomography, sound recordings, neural signaling, mechanical measurements, photography, videography, and

complex clinicopathological multivariate data. We have discussed a variety of applications in ENT along the main subspecialities from head and neck cancer, thyroid and parathyroid surgery, otology, rhinology, and laryngology. While some areas are focused on prediction in diagnosis and prognosis others, especially in otology, attempt to enhance established technologies. There is little doubt AI-related research will benefit otolaryngology in the near future with appropriate direction and judicious adoption.

Increasing literacy in data science and machine learning in otolaryngologists is a necessity for the progression of the field. It will help with clinical integration of algorithms and promote early adoption of AI tools once safe and effective. It will cultivate interest in this rapidly developing field and allow better cooperation with data scientists to direct clinically relevant research. Furthermore, it will facilitate the much-needed multicenter collaboration to take ENT AI research into the future. Although there are clear and acute challenges and limitations to this technology, AI developed systems will help define ENT and health care as a whole in the coming years.

References

- Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44–56.
- LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–44.
- Burt JR, Torosdagli N, Khosravan N, RaviPrakash H, Mortazi A, Tissavirasingham F, et al. Deep learning beyond cats and dogs: recent advances in diagnosing breast cancer with deep neural networks. *Br J Radiol*. 2018;91(1089):20170545.
- Liu Y, Logan B, Liu N, Xu Z, Tang J, Wang Y. Deep reinforcement learning for dynamic treatment regimes on medical registry data. *Healthc Inform*. 2017;2017:380–5.
- Liu S, See KC, Ngiam KY, Celi LA, Sun X, Feng M. Reinforcement learning for clinical decision support in critical care: comprehensive review. *J Med Internet Res*. 2020;22(7):e18477.
- Van Gerven M, Bohte S. Editorial: artificial neural networks as models of neural information processing. *Front Comput Neurosci*. 2017;11:114.
- McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577(7788):89–94.
- Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115–8.
- Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402–10.
- Peng Z, Wang Y, Wang Y, Jiang S, Fan R, Zhang H, et al. Application of radiomics and machine learning in head and neck cancers. *Int J Biol Sci*. 2021;17(2):475–86.
- Shen H, Wang Y, Liu D, Lv R, Huang Y, Peng C, et al. Predicting progression-free survival using MRI-based radiomics for patients with nonmetastatic nasopharyngeal carcinoma. *Front Oncol*. 2020;10:618.
- Liu Z, Cao Y, Diao W, Cheng Y, Jia Z, Peng X. Radiomics-based prediction of survival in patients with head and neck squamous cell carcinoma based on pre- and post-treatment (18)F-PET/CT. *Aging (Albany NY)*. 2020;12(14):14593–619.
- Ger RB, Zhou S, Elgohari B, Elhalawani H, Mackin DM, Meier JG, et al. Radiomics features of the primary tumor fail to improve prediction of overall survival in large cohorts of CT- and PET-imaged head and neck cancer patients. *PLoS One*. 2019;14(9):e0222509.
- Bologna M, Calareso G, Resteghini C, Sdao S, Montin E, Corino V, et al. Relevance of apparent diffusion coefficient features for a radiomics-based prediction of response to induction chemotherapy in sinonasal cancer. *NMR Biomed*. 2020;2020:e4265.
- Zhao L, Gong J, Xi Y, Xu M, Li C, Kang X, et al. MRI-based radiomics nomogram may predict the response to induction chemotherapy and survival in locally advanced nasopharyngeal carcinoma. *Eur Radiol*. 2020;30(1):537–46.
- Zhai TT, Langendijk JA, van Dijk LV, van der Schaaf A, Sommers L, Vemer-van den Hoek JGM, et al. Pre-treatment radiomic features predict individual lymph node failure for head and neck cancer patients. *Radiother Oncol*. 2020;146:58–65.
- Wu W, Ye J, Wang Q, Luo J, Xu S. CT-based radiomics signature for the preoperative discrimination between head and neck squamous cell carcinoma grades. *Front Oncol*. 2019;9:821.
- Mukherjee P, Cintra M, Huang C, Zhou M, Zhu S, Colevas AD, et al. CT-based radiomic signatures for predicting histopathologic features in head and neck squamous cell carcinoma. *Radiol Imaging Cancer*. 2020;2(3):e190039.
- Ren J, Qi M, Yuan Y, Tao X. Radiomics of apparent diffusion coefficient maps to predict histologic grade in squamous cell carcinoma of the oral tongue and floor of mouth: a preliminary study. *Acta Radiol*. 2020;62:453–61. <https://doi.org/10.1177/0284185120931683>.
- Wang F, Zhang B, Wu X, Liu L, Fang J, Chen Q, et al. Radiomic nomogram improves preoperative T

- category accuracy in locally advanced laryngeal carcinoma. *Front Oncol.* 2019;9:1064.
21. Romeo V, Cuocolo R, Ricciardi C, Ugga L, Cocozza S, Verde F, et al. Prediction of tumor grade and nodal status in oropharyngeal and oral cavity squamous-cell carcinoma using a radiomic approach. *Anticancer Res.* 2020;40(1):271–80.
22. Tomita H, Yamashiro T, Heianna J, Nakasone T, Kobayashi T, Mishiro S, et al. Deep learning for the preoperative diagnosis of metastatic cervical lymph nodes on contrast-enhanced computed tomography in patients with oral squamous cell carcinoma. *Cancers (Basel).* 2021;13(4):600.
23. Kann BH, Aneja S, Loganadane GV, Kelly JR, Smith SM, Decker RH, et al. Pretreatment identification of head and neck cancer nodal metastasis and extranodal extension using deep learning neural networks. *Sci Rep.* 2018;8(1):14036.
24. Jiang W, Lakshminarayanan P, Hui X, Han P, Cheng Z, Bowers M, et al. Machine learning methods uncover radiomorphologic dose patterns in salivary glands that predict xerostomia in patients with head and neck cancer. *Adv Radiat Oncol.* 2019;4(2):401–12.
25. Vrtovec T, Mocnik D, Strojan P, Pernus F, Ibragimov B. Auto-segmentation of organs at risk for head and neck radiotherapy planning: From atlas-based to deep learning methods. *Med Phys.* 2020;47(9):e929–e950.
26. Halicek M, Shahedi M, Little JV, Chen AY, Myers LL, Sumer BD, et al. Head and neck cancer detection in digitized whole-slide histology using convolutional neural networks. *Sci Rep.* 2019;9(1):14043.
27. Mascharak S, Baird BJ, Holsinger FC. Detecting oropharyngeal carcinoma using multispectral, narrow-band imaging and machine learning. *Laryngoscope.* 2018;128(11):2514–20.
28. Shenson JA, Liu GS, Farrell J, Blevins NH. Multispectral imaging for automated tissue identification of normal human surgical specimens. *Otolaryngol Head Neck Surg.* 2021;164(2):328–35.
29. Fei B, Lu G, Wang X, Zhang H, Little JV, Patel MR, et al. Label-free reflectance hyperspectral imaging for tumor margin assessment: a pilot study on surgical specimens of cancer patients. *J Biomed Opt.* 2017;22(8):1–7.
30. Stepp WH, Farquhar D, Sheth S, Mazul A, Mamdani M, Hackman TG, et al. RNA oncoimmune phenotyping of HPV-positive p16-positive oropharyngeal squamous cell carcinomas by nodal status. *JAMA Otolaryngol Head Neck Surg.* 2018;144(11):967–75.
31. Chang SW, Abdul-Kareem S, Merican AF, Zain RB. Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. *BMC Bioinformatics.* 2013;14:170.
32. Carnielli CM, Macedo CCS, De Rossi T, Granato DC, Rivera C, Domingues RR, et al. Combining discovery and targeted proteomics reveals a prognostic signature in oral cancer. *Nat Commun.* 2018;9(1):3598.
33. Bohnenberger H, Kaderali L, Strobel P, Yepes D, Plessmann U, Dharia NV, et al. Comparative proteomics reveals a diagnostic signature for pulmonary head-and-neck cancer metastasis. *EMBO Mol Med.* 2018;10(9):e8428.
34. Buda M, Wildman-Tobriner B, Hoang JK, Thayer D, Tessler FN, Middleton WD, et al. Management of thyroid nodules seen on US images: deep learning may match performance of radiologists. *Radiology.* 2019;292(3):695–701.
35. Thomas J, Haertling T. AIBx, artificial intelligence model to risk stratify thyroid nodules. *Thyroid.* 2020;30(6):878–84.
36. Wang H, Song B, Ye N, Ren J, Sun X, Dai Z, et al. Machine learning-based multiparametric MRI radiomics for predicting the aggressiveness of papillary thyroid carcinoma. *Eur J Radiol.* 2020;122:108755.
37. Wei R, Wang H, Wang L, Hu W, Sun X, Dai Z, et al. Radiomics based on multiparametric MRI for extra-thyroidal extension feature prediction in papillary thyroid cancer. *BMC Med Imaging.* 2021;21(1):20.
38. Mourad M, Moubayed S, Dezube A, Mourad Y, Park K, Torreblanca-Zanca A, et al. Machine learning and feature selection applied to SEER data to reliably assess thyroid cancer prognosis. *Sci Rep.* 2020;10(1):5176.
39. Elliott Range DD, Dov D, Kovalsky SZ, Henao R, Carin L, Cohen J. Application of a machine learning algorithm to predict malignancy in thyroid cytopathology. *Cancer Cytopathol.* 2020;128(4):287–95.
40. Akbulut S, Erten O, Kim YS, Gokceimam M, Berber E. Development of an algorithm for intraoperative autofluorescence assessment of parathyroid glands in primary hyperparathyroidism using artificial intelligence. *Surgery.* 2021;170:454.
41. Maktabi M, Kohler H, Ivanova M, Neumuth T, Rayes N, Seidemann L, et al. Classification of hyperspectral endocrine tissue images using support vector machines. *Int J Med Robot.* 2020;16(5):1–10.
42. Imbus JR, Randle RW, Pitt SC, Sippel RS, Schneider DF. Machine learning to identify multigland disease in primary hyperparathyroidism. *J Surg Res.* 2017;219:173–9.
43. You E, Lin V, Mijovic T, Eskander A, Crowson MG. Artificial intelligence applications in otology: a state of the art review. *Otolaryngol Head Neck Surg.* 2020;163(6):1123–33.
44. Fauser J, Stenin I, Bauer M, Hsu WH, Kristin J, Klenzner T, et al. Toward an automatic preoperative pipeline for image-guided temporal bone surgery. *Int J Comput Assist Radiol Surg.* 2019;14(6):967–76.
45. Wang YM, Li Y, Cheng YS, He ZY, Yang JM, Xu JH, et al. Deep learning in automated region proposal and diagnosis of chronic otitis media based on computed tomography. *Ear Hear.* 2020;41(3):669–77.
46. Burwood GWS, Fridberger A, Wang RK, Nuttall AL. Revealing the morphology and function of the cochlea and middle ear with optical coherence tomography. *Quant Imaging Med Surg.* 2019;9(5):858–81.

47. Liu GS, Zhu MH, Kim J, Raphael P, Applegate BE, Oghalai JS. ELHnet: a convolutional neural network for classifying cochlear endolymphatic hydrops imaged with optical coherence tomography. *Biomed Opt Express.* 2017;8(10):4579–94.
48. Abdollahi H, Mostafaei S, Cheraghi S, Shiri I, Rabi Mahdavi S, Kazemnejad A. Cochlea CT radiomics predicts chemoradiotherapy induced sensorineural hearing loss in head and neck cancer patients: A machine learning and multi-variable modelling study. *Phys Med.* 2018;45:192–7.
49. Oyewumi M, Brandt MG, Carrillo B, Atkinson A, Iglar K, Forte V, et al. Objective evaluation of otoscopy skills among family and community medicine, pediatric, and otolaryngology residents. *J Surg Educ.* 2016;73(1):129–35.
50. Lee JY, Choi S-H, Chung JW. Automated classification of the tympanic membrane using a convolutional neural network. *Appl Sci.* 2019;9(9):1827.
51. Wu Z, Lin Z, Li L, et al. Deep Learning for Classification of Pediatric Otitis Media. *Laryngoscope.* 2021;131(7):E2344–E2351.
52. McKearney RM, MacKinnon RC. Objective auditory brainstem response classification using machine learning. *Int J Audiol.* 2019;58(4):224–30.
53. Bing D, Ying J, Miao J, Lan L, Wang D, Zhao L, et al. Predicting the hearing outcome in sudden sensorineural hearing loss via machine learning models. *Clin Otolaryngol.* 2018;43(3):868–74.
54. Zhao Y, Li J, Zhang M, Lu Y, Xie H, Tian Y, et al. Machine learning models for the hearing impairment prediction in workers exposed to complex industrial noise: a pilot study. *Ear Hear.* 2019;40(3):690–9.
55. Kates JM, Arehart KH, Souza PE. Integrating cognitive and peripheral factors in predicting hearing-aid processing effectiveness. *J Acoust Soc Am.* 2013;134(6):4458.
56. Kumar A, Florencio D. Speech enhancement in multiple-noise conditions using deep neural networks. *arXiv pre-print server.* 2016.
57. Tahmasebi S, Gajecki T, Nogueira W. Design and evaluation of a real-time audio source separation algorithm to remix music for cochlear implant users. *Front Neurosci.* 2020;14:434.
58. Gartner L, Lenarz T, Joseph G, Buchner A. Clinical use of a system for the automated recording and analysis of electrically evoked compound action potentials (ECAPs) in cochlear implant patients. *Acta Otolaryngol.* 2010;130(6):724–32.
59. Sinkiewicz D, Friesen L, Ghoraani B. A novel method for extraction of neural response from single channel cochlear implant auditory evoked potentials. *Med Eng Phys.* 2017;40:47–55.
60. Pile J, Wanna GB, Simaan N. Robot-assisted perception augmentation for online detection of insertion failure during cochlear implant surgery. *Robotica.* 2017;35(7):1598–615.
61. Exarchos TP, Rigas G, Bibas A, Kikidis D, Nikitas C, Wuyts FL, et al. Mining balance disorders' data for the development of diagnostic decision support systems. *Comput Biol Med.* 2016;77:240–8.
62. Krafczyk S, Tietze S, Swoboda W, Valkovic P, Brandt T. Artificial neural network: a new diagnostic posturographic tool for disorders of stance. *Clin Neurophysiol.* 2006;117(8):1692–8.
63. Chowdhury NI, Smith TL, Chandra RK, Turner JH. Automated classification of osteomeatal complex inflammation on computed tomography using convolutional neural networks. *Int Forum Allergy Rhinol.* 2019;9(1):46–52.
64. Huang J, Habib AR, Mendis D, Chong J, Smith M, Duvnjak M, et al. An artificial intelligence algorithm that differentiates anterior ethmoidal artery location on sinus computed tomography scans. *J Laryngol Otol.* 2020;134(1):52–5.
65. Parmar P, Habib AR, Mendis D, Daniel A, Duvnjak M, Ho J, et al. An artificial intelligence algorithm that identifies middle turbinate pneumatisation (concha bullosa) on sinus computed tomography scans. *J Laryngol Otol.* 2020;134(4):328–31.
66. Deutschmann MW, Yeung J, Bosch M, Lysack JT, Kingstone M, Kilty SJ, et al. Radiologic reporting for paranasal sinus computed tomography: a multi-institutional review of content and consistency. *Laryngoscope.* 2013;123(5):1100–5.
67. Wu Q, Chen J, Deng H, Ren Y, Sun Y, Wang W, et al. Expert-level diagnosis of nasal polyps using deep learning on whole-slide imaging. *J Allergy Clin Immunol.* 2020;145(2):698–701.e6.
68. Dimauro G, Ciprandi G, Deperte F, Girardi F, Ladisa E, Latrofa S, et al. Nasal cytology with deep learning techniques. *Int J Med Inform.* 2019;122:13–9.
69. Fokkens WJ, Lund VJ, Hopkins C, Hellings PW, Kern R, Reitsma S, et al. European position paper on rhinosinusitis and nasal polyps 2020. *Rhinology.* 2020;58(Suppl S29):1–464.
70. Soler ZM, Hyer JM, Rudmik L, Ramakrishnan V, Smith TL, Schlosser RJ. Cluster analysis and prediction of treatment outcomes for chronic rhinosinusitis. *J Allergy Clin Immunol.* 2016;137(4):1054–62.
71. Liao B, Liu JX, Li ZY, Zhen Z, Cao PP, Yao Y, et al. Multidimensional endotypes of chronic rhinosinusitis and their association with treatment outcomes. *Allergy.* 2018;73(7):1459–69.
72. Schonweiler R, Hess M, Wubbelt P, Ptak M. Novel approach to acoustical voice analysis using artificial neural networks. *J Assoc Res Otolaryngol.* 2000;1(4):270–82.
73. Fang SH, Tsao Y, Hsiao MJ, Chen JY, Lai YH, Lin FC, et al. Detection of pathological voice using cepstrum vectors: a deep learning approach. *J Voice.* 2019;33(5):634–41.
74. Cesari U, De Pietro G, Marciano E, Niri C, Sannino G, Verde L. Voice disorder detection via an m-health system: design and results of a clinical study to evaluate Vox4Health. *Biomed Res Int.* 2018;2018:8193694.
75. Unger J, Lohscheller J, Reiter M, Eder K, Betz CS, Schuster M. A noninvasive procedure for early-stage

- discrimination of malignant and precancerous vocal fold lesions based on laryngeal dynamics analysis. *Cancer Res.* 2015;75(1):31–9.
76. Tsui SY, Tsao Y, Lin CW, Fang SH, Lin FC, Wang CT. Demographic and symptomatic features of voice disorders and their potential application in classification using machine learning algorithms. *Folia Phoniatr Logop.* 2018;70(3–4):174–82.
77. Wang TV, Adamian N, Song PC, Franco RA, Huston MN, Jowett N, et al. Application of a computer vision tool for automated glottic tracking to vocal fold paralysis patients. *Otolaryngol Head Neck Surg.* 2021;194:599821989608.
78. Zhang Z, Coyle JL, Sejdic E. Automatic hyoid bone detection in fluoroscopic images using deep learning. *Sci Rep.* 2018;8(1):12310.
79. Donohue C, Mao S, Sejdić E, Coyle JL. Tracking Hyoid Bone Displacement During Swallowing Without Videofluoroscopy Using Machine Learning of Vibratory Signals. *Dysphagia.* 2021;36(2):259–269.
80. Donohue C, Khalifa Y, Perera S, Sejdić E, Coyle JL. How Closely do Machine Ratings of Duration of UES Opening During Videofluoroscopy Approximate Clinician Ratings Using Temporal Kinematic Analyses and the MBSImP?. *Dysphagia.* 2021;36(4):707–718.
81. Khalifa Y, Donohue C, Coyle JL, Sejdic E. Upper esophageal sphincter opening segmentation with convolutional recurrent neural networks in high resolution cervical auscultation. *IEEE J Biomed Health Inform.* 2021;25(2):493–503.
82. Organization WH. Big data and artificial intelligence 2019. <https://www.who.int/ethics/topics/big-data-artificial-intelligence/en/>
83. Colleges AoRM. Artificial Intelligence in Healthcare 2019 01.03.2021. https://www.aomrc.org.uk/wp-content/uploads/2019/01/Artificial_intelligence_in_healthcare_0119.pdf
84. Howard J. Artificial intelligence: implications for the future of work – CDC 2019. <https://blogs.cdc.gov/niosh-science-blog/2019/08/26/ai/>
85. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise QC. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak.* 2020;20(1):310.
86. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. *JAMA Dermatol.* 2018;154(11):1247–8.
87. Narla A, Kuprel B, Sarin K, Novoa R, Ko J. Automated classification of skin lesions: from pixels to practice. *J Invest Dermatol.* 2018;138(10):2108–10.
88. Oakden-Rayner L. Exploring the ChestXray14 dataset: problems. 2017. <https://lukeoakdenrayner.wordpress.com/2017/12/18/>



Shravanti Muthu, Fatima Nabi, and Junaid Nabi

Contents

Introduction	1003
AI in IVF	1004
Ethical Challenges	1005
References	1005

Abstract

With persistently low fertility rates at the global levels, combined with the anticipated capacitive insufficiency in specialist clinicians – in both developed and developing countries – the need for addressing the concerns of patients who intend to conceive is greater than ever. In an effort to advance the treatment of patients who experience infertility, AI is increasingly being used to optimize the clinical management options. The major areas of application continue to be in the space of analyzing semen, assessment of oocytes and embryos, and predictive analytics for the success of in vitro fertilization. Given how these automated decision systems

often operate as a “black box,” with both the technical developers and clinical users often unaware of the potential outcomes, ethical concerns around the limits of this technology remain.

Keywords

Artificial intelligence · Artificial intelligence in obstetrics · Artificial intelligence in genecology · Machine learning algorithms · Precision oncology · Automated decision-making systems · Supervised learning · Training sets in obstetrics and genecology

Introduction

When coining the term “artificial intelligence” – by John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon at a Dartmouth conference in late summer of 1955 – it was rather inconceivable the extent of impact and the range of applications that this technology would demonstrate [1]. Now, artificial intelligence (AI) is

S. Muthu
Boston University School of Public Health, Boston, MA,
USA

F. Nabi
Boston, MA, USA

J. Nabi (✉)
Harvard University, Boston, MA, USA

poised to change the delivery of healthcare as clinicians know it. The applications – and the ensuing clinical trials – that apply AI to augment the delivery of services such as pathology and radiology are clear; however, there is another area where AI is bound to play a critical role in addressing an unmet need: obstetrics and gynecology.

AI can augment clinicians by assisting with earlier diagnoses, changing the trajectory for the patient in potentially fatal conditions – such as ovarian and cervical cancers. It can sift through data much faster and more efficiently than human capacity, comparing target sample with a general sample in order to identify tumor markers, irregular genetic code, and presence of abnormal proteins. With prompt diagnosis, clinicians can not only curb case fatalities but also reduce physician burden and cost in eliminating ineffective treatments all the while reducing human suffering.

Now, with persistently low fertility rates at the global levels [2], combined with the anticipated capacitive insufficiency in specialist clinicians – in both developed and developing countries [3] – the need for addressing the concerns of patients who intend to conceive is greater than ever. In this chapter, we underscore the applicability of AI for in vitro fertilization (IVF) and bioethical issues around utilizing this technology.

AI in IVF

Evidence suggests that the quality of the embryo is often a determinant for IVF success; at the same time, technologies that will enable an accurate assessment of sperms, oocytes, as well as embryos are often not robust. In an effort to advance the treatment for patients who experience infertility, AI is increasingly being used to optimize the clinical management options. The major areas where AI is applicable in the field of obstetrics, especially in vitro fertilization, are analyzing semen, assessment of oocytes and embryos, and predictive analytics for the success of in vitro fertilization.

Analysis of semen is often the initial diagnostic tool available for infertile couples. Identifying

abnormalities with the sperm cells – especially morphology and motility – can be a significant determinant of eventual fertility. Currently, computer-aided sperm analysis (CASA) systems can assess these characteristics in sperm cells. Using AI-enabled image analysis has the potential to lead to results that are more precise and reduce variability found in the manual inspection. In 2017, researchers of the University of North Carolina at Chapel Hill applied an automated technique to categorize motility patterns of human sperm – and the model accuracy was 89.92% [4].

Another area of promise for AI in the IVF space is that of assessment of oocytes and embryos. Given that the current rate of pregnancy occurrence for every viable oocyte remains under 5% [5], the demand for automated decision systems that can address lack of options for determining success of the IVF episode by the quality of oocyte is high [6]. Often, the manual inspection of oocytes can lead to misleading results – with cells visually appearing within routine morphological characteristics – leading to chromosomal abnormalities in the fetus [7]. This constitutes one of the most compelling reasons for opportunities to employ computational intelligence for such cases. Similarly, the ability of the embryo to survive is a crucial component for increasing the rate of pregnancy and the success of IVF approach [8]. Contemporary techniques employed – including visual inspection of the embryo morphology – are plagued by specialist variability that is closely related with the training of the specialist rather than the accuracy of their methods [9–11]. These concerns around embryo viability are more serious in the IVF approach, as increased repetition of embryos being placed from in vitro to in vivo can lead to complications for the expectant mother [12].

Although IVF remains an important option for infertile couples, the high costs – and often high rates of failure – make the approach prohibitive for many [13]. Automating parts of the process – and developing an ability to predict the success of the procedure with a convincing level of confidence – presents a highly sought-after use case for implementing AI in obstetrics. Early evidence suggests that predictive analytics in estimating

embryo viability through AI-enabled technologies hold promise – with recent studies reporting differentiation capability being up to 80% [14].

Ethical Challenges

There are some clear advantages of employing AI for IVF – potential to improve assessment of gametes, reduce manual inspection, and alleviate errors. However, as is the case with numerous AI-enabled medical technologies, application of AI methods in the space of in vitro fertilization or other gynecological diseases presents severe ethical dilemmas. For instance, the moral repercussions or legal liability concerns are not clear if embryos chosen by the automated decision-making technologies are deemed to have genetic abnormalities or chromosomal aberrations, especially after implantation or during the course of the pregnancy. Given how these automated decision systems operate as a “black box,” with both the technical developers and clinical users often unaware of the potential outcomes that do not always adhere to human understanding [15], the basis of assessments for the quality of a particular sperm, oocyte, or embryo remains relatively unknown. Another severe ethical concern is that of the definition of a “normal” embryo and the potential selection bias in data that determines this status [16]. This could lead to the unscrupulousness of designating certain genes or features as desirable at the expense of other features – and we know that sociocultural norms play a significant role in such determinations [17].

References

- Kreatsoulas C, Subramanian SV. Machine learning in social epidemiology: learning from experience. *SSM Popul Health.* 2018;4:347–9.
- Skakkebaek NE, Rajpert-De Meyts E, Buck Louis GM, Toppari J, Andersson AM, Eisenberg ML, et al. Male reproductive disorders and fertility trends: influences of environment and genetic susceptibility. *Physiol Rev.* 2016;96(1):55–97.
- Yang W, Williams JH, Hogan PF, Bruinooge SS, Rodriguez GI, Kosty MP, et al. Projected supply of and demand for oncologists and radiation oncologists through 2025: an aging, better-insured population will result in shortage. *J Oncol Pract.* 2014;10(1):39–45.
- Goodson SG, White S, Stevans AM, Bhat S, Kao CY, Jaworski S, et al. CASanova: a multiclass support vector machine model for the classification of human sperm motility patterns. *Biol Reprod.* 2017;97(5):698–708.
- Stoop D, Ermini B, Polyzos NP, Haentjens P, De Vos M, Verheyen G, et al. Reproductive potential of a metaphase II oocyte retrieved after ovarian stimulation: an analysis of 23 354 ICSI cycles. *Hum Reprod.* 2012;27(7):2030–5.
- Conti M, Franciosi F. Acquisition of oocyte competence to develop as an embryo: integrated nuclear and cytoplasmic events. *Hum Reprod Update.* 2018;24(3): 245–66.
- Munné S, Chen S, Colls P, Garrisi J, Zheng X, Cekleniak N, et al. Maternal age, morphology, development and chromosome abnormalities in over 6000 cleavage-stage embryos. *Reprod Biomed Online.* 2007;14(5):628–34.
- Saeedi P, Yee D, Au J, Havelock J. Automatic identification of human blastocyst components via texture. *IEEE Trans Biomed Eng.* 2017;64(12):2968–78.
- Baxter Bendus AE, Mayer JF, Shipley SK, Catherino WH. Interobserver and intraobserver variation in day 3 embryo grading. *Fertil Steril.* 2006;86(6):1608–15.
- Manna C, Nanni L, Lumini A, Pappalardo S. Artificial intelligence techniques for embryo and oocyte classification. *Reprod Biomed Online.* 2013;26(1):42–9.
- Santos Filho E, Noble JA, Poli M, Griffiths T, Emerson G, Wells D. A method for semi-automatic grading of human blastocyst microscope images. *Hum Reprod.* 2012;27(9):2641–8.
- Bromer JG, Seli E. Assessment of embryo viability in assisted reproductive technology: shortcomings of current approaches and the emerging role of metabolomics. *Curr Opin Obstet Gynecol.* 2008;20(3): 234–41.
- De Geyter C, Calhaz-Jorge C, Kupka MS, Wyns C, Mocanu E, Motrenko T, et al. ART in Europe, 2014: results generated from European registries by ESHRE: The European IVF-monitoring Consortium (EIM) for the European Society of Human Reproduction and Embryology (ESHRE). *Hum Reprod.* 2018;33(9):1586–601.
- Hafiz P, Nematollahi M, Boostani R, Namavar JB. Predicting implantation outcome of in vitro fertilization and intracytoplasmic sperm injection using data mining techniques. *Int J Fertil Steril.* 2017;11(3): 184–90.
- Nabi J. Addressing the “Wicked” problems in machine learning applications – time for bioethical agility. *Am J Bioeth.* 2020;20(11):25–7.
- Senders JT, Zaki MM, Karhade AV, Chang B, Gormley WB, Broekman ML, et al. An introduction and overview of machine learning in neurosurgical care. *Acta Neurochir (Wien).* 2018;160(1):29–38.
- Nabi J. How bioethics can shape artificial intelligence and machine learning. *Hastings Cent Rep.* 2018;48(5): 10–3.



Charles L. Bormann and Carol Lynn Curchoe

Contents

Introduction	1008
Artificial Intelligence in Reproductive Medicine	1008
Opportunities and Limitations of AI in Reproductive Medicine	1008
AI for Assessment, Diagnosis, or Treatment of Infertility	1011
AI for Embryo Annotation, Evaluation, and Selection	1011
AI for Prediction of Embryo Chromosome Status (Ploidy)	1012
AI in Maternal Healthcare Miscarriage Prediction	1012
Conclusion	1013
References	1013

Abstract

Artificial intelligence (AI) systems have been proposed for reproductive medicine since 1997. Artificial intelligence (AI) is perfectly suited to solve some of the complex problems in reproduction and pregnancy including assessment and diagnosis of infertility; oocyte and sperm analysis and selection; embryo annotation, evaluation, and selection; prediction of embryo chromosome status (ploidy); pregnancy viability; and miscarriage prediction. AI uses

multiple sources of data to reveal patterns in diagnosis, treatment, and results. AI systems are being developed to predict individual infertility patient risk, suggest treatment options in reproductive medicine, assess the chance of achieving a pregnancy and a healthy baby, anticipate complications during pregnancy, and predict “time to pregnancy.” With respect to IVF lab efficiency, manual processes and procedures currently predominate in the IVF laboratory. Automation and AI systems promise to lessen the burden of subjective, menial, or mundane aspects of the embryology laboratory, while automation can decrease inter- and intra-technician variability. The future of AI technologies promises to further address environmental stressors that can impair gamete function and embryo development.

C. L. Bormann
Massachusetts General Hospital IVF Laboratory, Boston,
MA, USA
e-mail: cbormann@partner.org

C. L. Curchoe (✉)
Fertility Guidance Technologies, Newport Beach, CA,
USA
e-mail: carol@fertilityguidancetechnologies.com

Keywords

Artificial intelligence · Machine learning · Artificial neural networks · Convolutional neural networks · Deep learning · IVF · Embryology · Assisted reproductive technology · ART · ICSI · Embryology · Andrology

Introduction

The applications of artificial intelligence in obstetrics and gynecology have grown significantly in the last 15 years. The health of woman during pregnancy can be effectively monitored and the provision of services improved. Disorders such as congenital heart birth defects or macrosomia, gestational diabetes, and preterm birth can be detected earlier when artificial intelligence is used. Infertility is another significant reproductive arena that AI is now significantly being developed and validated for. Millions of babies have been born through assisted reproductive technologies (ART); however, only 30% of in vitro fertilization (IVF) cycles succeed in a clinical pregnancy. Aside from increasing the success rate, there are other worthwhile goals for continued improvement across the industry, to simply get patients pregnant faster, reduce treatment dropout, or to reduce embryo wastage. Innovations in artificial intelligence (AI) will drive ART that is more reproducible, standardized, efficient, and less costly.

Artificial Intelligence in Reproductive Medicine

Opportunities and Limitations of AI in Reproductive Medicine

Predictive algorithms using AI have gained particular prominence in the field of reproduction (Table 1), maternal health, and assisted reproductive technology. The goals of automation through artificial intelligence are to improve efficacy, efficiency, and consistency of clinical decision

Table 1 Artificial intelligence and machine learning for maternal health

Preconception	Reproductive urology Genetic compatibility (HLA) Fertility management Family planning
Assisted reproductive technologies	Ovarian stimulation Gamete selection and grading Embryo grading Embryo selection Embryo ploidy Risk assessment
Prenatal care	Fetal anomalies Fetal size Fetal heart rate Ectopic pregnancy Placental function Miscarriage Pharmacological safety Gestational disease Maternal mortality
Preterm birth	Fetal mortality Fetal prognosis Contraction pattern Cervical properties Perinatal outcome prediction
Perinatal care/birth	Perinatal complication and APGAR score Date of delivery Complications
Postnatal/postpartum care	Postnatal health Postpartum depression Mortality

making during an ART treatment, such as; embryo classification and selection, sperm identification and classification, and ovarian response prediction and personalized protocols. Predictive modeling and artificial intelligence applications have evolved into a standalone subdiscipline of reproductive medicine, and the literature have been analyzed and evaluated using formal systems, such as the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). Predictive modeling and AI's place as a support tool is growing in clinical practice, but gains in patient outcomes and improved workflows have yet to be fully realized.

The development of AI systems that are widely applicable across ART clinics and protocols, and to geographies and populations, has been

challenged by the quality, diversity, and volume of available data. AI systems need big data to achieve robustness. Accessing so-called “BIG” data for ART should be a trivial problem, as more than 2.7 million assisted reproduction cycles are performed each year. However, it is a challenge for researchers to access “Big Data” and collaboration for prospective, cohort studies. These data are highly sensitive and strictly regulated, and patient consent and ethical approval are mandatory.

Large datasets are becoming increasingly important. So-called “big” data is used for research, external quality assurance, and policy making. Over 2.7 million assisted reproductive technology (ART) cycles are performed annually, generating millions upon millions of individual data points. ART is purportedly one of the best registered procedures in medicine, with global IVF registries spanning Europe, USA, Latina America, Japan, Africa, Canada, New Zealand, and Australia [10]. The International Committee Monitoring ART (ICMART) has summarized global IVF data [14] based on existing registries from approximately 60 countries and 2,500 centers (collectively representing approximately 4.5 million IVF cycles) spanning 2008–2010. Unfortunately, many limitations exist, making these databases moderately useful. Globally, ICMART is only capable of collecting unverifiable summary data. At the national level, datasets such as the HFEA were established for regulatory, not research purposes and therefore, lack the detail necessary to answer basic clinical practice questions.

Additionally, huge data sinkholes exist. For example, the majority of IVF lab data are recorded on paper charts, many labs don’t record any image data, outcome metrics are extremely variable, and the Asia Pacific region (where an estimated 400,000 (plus) cycles are performed annually) does not (yet) report outcome data to a national registry.

Another challenge faced in AI development is the quality of the training data and how the machines learn from it. The training data – perhaps embryo, oocyte, or sperm images – still have to be evaluated and graded by human

embryologists, and the populations these training data come from may be biased (by income, race, infertility diagnosis, etc.). How well the AI learns depends on the quality and size of the dataset. The role of medical coders has not yet been fully realized for the development of AI systems for ART. Medical coders analyze individual patient charts and translate complex information about diagnoses, treatments, medications, and more into alphanumeric codes. These codes are submitted to billing systems and health insurers for payment and reimbursement and play a critical role in patient care. Billing specialists represent an untapped well of knowledge; as they are responsible for the accurate assessment of charts, they could also enrich AI systems with medical knowledge that can improve the system’s performance.

A critical step in modeling is preprocessing and the treatment of images, before feature extraction and model training. Clumsy preprocessing and federated subsets of data generated at various geographic locations (with different cropping and contrast, taken on a wide variety of cameras and microscopes) may yield an inconsistent database that leads AI models to poor accuracy. A worldwide common reproduction specific git style image and video repository (similar to genome sequence databases) and automated methods to crop, remove patient information, rotate and size images or video is a worthwhile goal, and one that will take immense international collaboration. Commercial blockchain solutions are already being built, to provide distributed frameworks to enable privacy preservation, collaborative machine learning, and guaranteed traceability and authenticity of data using a distributed ledger.

Typically, AI systems are trained with many billions or millions of data points. As of the writing of this chapter, the largest published datasets that have been used to train or validate various deep learning, machine learning, and neural network systems in the field of reproduction fall in the 1,000–10,000 range, for example, the following:

- Bormann, C. et al. [1] (742 embryo images)
Kan-Tor et al. [2] (6,200 time lapse videos); 52 Chavez-Badiola, A. [33] (1,231 embryo images)

VerMilyea M. et al. [3] (8,886 embryos)
Letterie, G. et al. [17] (2,603 total cycles (1,853 autologous and 750 donor cycles))
Tran, D. et al. [4] (8,836 embryos)
Blank, C. et al. [5] (1,052 patients)
Chen, T et al. [6] (171,239 images from 16,201 embryos)
Miyagi, Y. et al. [20] (5,691 embryo images)

AI for reproduction dataset is available for download at www.repro-ai.org.

It is likely that the majority of AI training data are generated mainly from IVF cycles in a very specific patient population – white, wealthy, etc. There is a marked, and unacceptable, difference in the outcomes after fertility treatment between minority populations. The assisted reproductive technology (ART) failure rate, defined as no live birth after treatment, is 51.9% (white), 61.8% (Asian), 62.2% (Black), and 55.9% (Hispanic), and ART stillbirths account for 16.3% (white), 18.4% (Asian), 25.0% (Black), and 17.8% (Hispanic).

These are problems that AI is uniquely suited to solve. Artificial intelligence in “precision reproductive medicine” has the potential to resolve the genetic uniqueness of individuals and the molecular mechanisms of their infertility, to match each individual to the best treatment.

Conventional IVF (despite scientific advances like genetic carrier screening, embryo biopsy with pre-implantation genetic diagnosis, ICSI, and “freeze-all” cycles) is woefully inefficient and wastes millions of gametes and embryos worldwide. Convolutional neural networks have been shown to choose genetically “normal” blastocysts from single static images or video, better than fully trained embryologists with dozens of years of experience. Advances like these, when applied clinically, have the potential to address this problem.

In the USA, there is a vast underserved infertility market that includes those seeking genetic disease prevention, LGBTQ, proactive fertility management (oocyte and sperm freezing), and those without access to fertility care (i.e., most people, as only 17 of the USA require infertility insurance, although private employers are

increasingly offering these benefits). We know beyond a shadow of a doubt this is true, by comparison with other developed countries. There are seven million women with infertility in the USA, and only 1% of those successfully deliver an in vitro fertilization baby per year. Australia, Japan, and in many parts of Europe have rates that are three times higher than the USA. IVF is an incredibly manual labor-intensive process.

There is a dearth of well-trained embryologists to perform these labor-intensive tasks, and staff training is often a “bottleneck” in the management of IVF labs. AI systems have the potential to revolutionize the market by assisting with the manual, time and labor consuming, rote, routine, intensive and subjective embryology tasks while increasing the number of patients (scale) that can be served safely.

There are significant challenges to implementing any AI system in a meaningful way into a clinical IVF laboratory, yet it is almost certain that AI is the future of reproductive medicine. Wherever electronic health records (generally), and AI systems specifically, have been proposed, there are unresolved controversies about privacy, effects on doctor/patient relationships and trust, whether individuals can opt out, who controls or can have access to data and when consent is required, how to protect data from unauthorized use or accidental loss, and use of “de-identified” data for ethically approved research.

Implementation of an AI at the actual point of care in a routine, easy, and automated fashion is also currently impossible for the majority of ART clinics that still use paper charts and doesn’t capture and store even a single digital image of their patient’s embryos or gametes, much less the videos that are needed for true big data. Additionally, data management solutions to augment patient demographics, clinical and lab key performance indicators (KPI), and other relevant data streams (ultrasound images and PGT results, competency assessments, endocrinology, microscope and incubator QC, room/environmental factors) into a single dashboard are lacking.

Research laboratories worldwide have developed AI systems to enhance diagnostic and prognostic capabilities, minimize variability and error,

and maximize precision and efficiency of IVF treatment, but they have not yet achieved robustness for safe clinical implementation, and no prospective validation trials have been published as of the writing of this chapter.

Looking to the future, research scientists and clinicians are tasked to transcend simply packaging *established* pregnancy predictors in fancy new machine learning algorithms. *Novel* variables and *hidden* relationships that allow for superior predictive capabilities will be uncovered through prospective design, big datasets, standardized outcome measures, external validation, and integration of clinical and laboratory key performance indicators, and patient demographics.

Current controversies in the field illuminate the difficulties with subjective grading, false negatives, mosaic embryos, and the near certainty that viable embryos are being discarded even using the best techniques available. Bias and black-box problems mean that AI systems that seek to diagnose, rather than rank embryos for transfer have the potential to worsen this problem.

AI for Assessment, Diagnosis, or Treatment of Infertility

Infertility is a multifactor disease, which makes diagnosis and treatment complicated. In reproductive urology, AI applications initially focused on semen parameters [15] but have advanced to automated sperm detection, semen analysis, and prediction of disease [29], for example, in the patient population most likely to need a genetic workup for azoospermia [19]. Goodson et al. [30] classified individual sperm as progressive, intermediate, hyperactivated, slow, or weakly motile using support vector machines (SVM) with 89.9% accuracy. Mirsky et al. [31] employ interferometric phase microscopy along with SVM to develop a model that assess the morphology of sperms and classify them into good or bad morphology with over 88% accuracy. Girela et al. [32] aimed to evaluate the relationship of lifestyle habits and semen quality in young healthy men, using 34 environment and lifestyle variables. They

were able to determine sperm concentration with 90% accuracy and sperm motility with 82% accuracy, while El-Shafeiy et al. [7] found that just nine lifestyle parameters could predict semen quality.

Liao et al. [34] showed that a machine learning-derived algorithm is useful to assist clinicians in making an efficient and accurate initial judgment on the condition of patients with infertility. Over 60,000 infertile couples' medical records were evaluated using a grading system that divided the condition of the patients into 5 grades from A to E. The worst E grade represented a 0.90% pregnancy rate, and the pregnancy rate in the A grade was 53.82%. The cross-validation results showed that the stability of the system was 95.94%.

Letterie et al. [17] evaluated a computer decision support system for a day-to-day management of ovarian stimulation during in vitro fertilization following key decisions made during an IVF cycle: (1) stop stimulation or continue stimulation. If the decision was to stop, then the next automated decision was to (2) trigger or cancel. If the decision was to continue stimulation, then the next key decisions were (3) number of days to follow-up and (4) whether any dose adjustment was needed. They used data derived from an electronic medical records system of a female population undergoing IVF cycles and oocyte cryopreservation to include the patient's demographics, past medical history, and infertility evaluation including diagnosis, laboratory testing for ovarian reserve, and any radiologic studies pertinent to a diagnosis of infertility. The four key decisions during the process of ovarian stimulation and IVF were compared to expert decisions across 12 providers and found to have a sensitivity of 0.98 for trigger and 0.78 for cycle cancellation. Nisal et al. show a 40% decrease in stimulation medications, as suggested by AI systems [24].

AI for Embryo Annotation, Evaluation, and Selection

It is generally acknowledged that despite advances in embryo selection based on morphology, time-

lapse photography, or embryo biopsy with preimplantation genetic testing (PGT), implantation rates in human IVF are incredibly inefficient. While morphological evaluation is widely accepted and implemented in most IVF clinics worldwide, the exact morphologic parameters used to score embryos are highly variable between clinics, and the assessment process is strikingly subjective even between embryologists at the same clinic.

The development of automatic time-lapse imaging (TLI) systems (the unification of an incubator and microscope connected to a computer for automated image capture and recording) has made it possible for embryologists to continuously monitor embryonic development without disturbing *in vitro* culture conditions. In the decade since it was introduced to the ART field, TLI has enabled embryologists and researchers to thoroughly track and annotate the specific timing and duration of morphological changes, known today as morphokinetics, as well as identify previously unobserved embryonic events. This new source of data has paved the way for an improved understanding of the clinical implications of various morphokinetic attributes, and has made it possible to conduct detailed retrospective studies on embryos with a known clinical outcomes.

TLI systems are expensive and are not readily available to most fertility centers. Therefore, assistive tools and algorithms that can work with static (2D) images acquired from traditional microscopes that are available to virtually all fertility centers are being developed. A multicenter study using large retrospective data of single blastocyst (BL) images used an artificial intelligence (AI) platform to predict FH implantation [3]. The model's predictions were compared to those made by embryologists across 11 different IVF labs and showed an overall improvement. Chavez-Badiola et al. [8] predict embryo ploidy using computer vision for feature extraction and image preprocessing on a single blastocyst image. Bormann et al. [1] examined the performance of a deep convolutional neural network (CNN), trained using embryo images from a single timepoint to detect which embryos were most likely to implant out of a group of high-quality embryos that displayed few visible differences.

AI for Prediction of Embryo Chromosome Status (Ploidy)

Embryo morphology (blastocyst grade and score) has been shown to be associated with embryo ploidy. High-quality embryos have a higher chance of being euploid. Noninvasive prediction of embryo ploidy is becoming increasingly important. Trophectoderm biopsy, performed at the blastocyst stage, is invasive and likely impairs embryo development, reducing the embryo development potential. Other downsides of trophectoderm biopsy are the cost of sequencing, the clinical management of mosaic embryos, the skill required to biopsy TE cells, and the fact that only a select number of blastocysts (BLs) can be tested. Noninvasive biopsy methods promise to provide comprehensive embryo information, prioritize embryos for trophectoderm biopsy, and successfully rank embryos for transfer.

AI in Maternal Healthcare Miscarriage Prediction

Several diseases and adverse outcomes during pregnancy (e.g., preterm birth, preeclampsia, and miscarriage) have complicated and difficult-to-understand etiology. In places like the USA, maternal mortality is increasing, unlike other places worldwide where it is decreasing. Pregnancy and maternal healthcare generate many different data types (e.g., ultrasound imaging, diagnostic screening, fetal monitoring, genetics) that can be integrated by AI to address maternal health throughout the pregnancy process: preconception, prenatal, perinatal, and postnatal period. There is a significant gap in the research on pregnant and lactating, breastfeeding women and pharmacological safety and efficacy due to their systemic exclusion from RCTs, yet up to 80% of pregnant or lactating women will need to take a pharmacological substance [16].

Ultrasound (US) images and videos are a significant source of potential data for AI analysis. Detecting or predicting miscarriage and pregnancy loss is of particular importance, particularly following infertility treatment when the patient

will already be closely monitored [18]. Among the ultrasound scan markers, fetal bradycardia is a strong marker for miscarriage prediction and could be combined with other markers such as intrauterine hematoma, crown rump length, yolk sac, and patient demographic and biochemical markers [13].

Another approach [21] explored the use of AI in combination with immunogenetic data (HLA) to predict recurrent MCs. This research implies that accurately assessing the risk of recurrent miscarriage associated with a given pair of gametes could improve gamete donor selection and therefore increase pregnancy success rates.

Paydar et al. developed a clinical decision support system to predict pregnancy outcomes among systemic lupus erythematosus (SLE)-affected pregnant women, namely, spontaneous abortion or live birth. Two ANNs were trained based on features selected by a binary logistic accurate (91%) for prediction of spontaneous abortion or live birth of SLE-affected pregnancy.

Several studies aim to predict and classify disease in early pregnancy, improve screenings, and provide clinical decision support for disease management. Chronic hypertension, chronic respiratory disease, substance-use disorders, and preexisting diabetes are common disorders that can coincide in pregnant patients. Pregnancy itself is the source of gestational diabetes mellitus (GDM), hypertension disorders, and bacteriuria (hemolysis, elevated liver enzymes, and low platelet (HELLP) syndrome). Polak and Mendyk [26] developed a GDM screening tool using ANNs to model relationships between demographic factors and the risk of GDM, while Moreira et al. [22] propose application of the radial basis function network (RBFNetwork), an ANN technique, to identify possible cases of GDM in pregnant women. Tejera et al. [27] constructed a model for classification of women with normal, hypertensive, and preeclamptic pregnancy using maternal heart rate variability indexes and ANNs. Moreira et al. [23] propose a model using ANNs and fuzzy logic to predict HELLP syndrome in high-risk pregnancies.

Boland et al. [12] developed a method that utilizes machine learning to predict the *fetal*

toxicity of pharmacologics taken during pregnancy, including first through third trimesters of the pregnancy.

Conclusion

Currently, prospective studies are necessary to validate any AI system's usefulness in the clinical setting of obstetrics and gynecology, urology, and infertility. In reproductive medicine, data is not as easily obtained due to factors like data privacy, and lack of structured EMRs among others. To overcome this problem, multicentric collaboration is one way in which some groups have managed to increase the size of their data with excellent results. Low-quality and/or small-sized databases can result in biased models, which might not be generalizable across clinics nor reproducible. Since AI relies on the quality of the information used for training, transparency about the quality of datasets becomes paramount to understanding a novel AI's clinical potential.

References

1. Bormann CL, Kanakasabapathy MK, Thirumalaraju P, Gupta R, Pooniwala R, Kandula H, Hariton E, Souter I, Dimitriadis I, Ramirez LB, Curchoe CL, Swain JE, Boehnlein LM, Shafiee H. Performance of a deep learning based neural network in the selection of human blastocysts for implantation. *elife*. 2020;9: e55301. <https://doi.org/10.7554/elife.55301>.
2. Kan-Tor Y, Zabari N, Erlich I, Szeskin A, Amitai T, Richter D, Or Y, Shoham Z, Hurwitz A, Har-Vardi I, Gavish M, Ben-Meir A, Buxboim A. Automated evaluation of human embryo blastulation and implantation potential using deep-learning. *Adv Intell Syst*. 2020;2 (10):2000080.
3. VerMilyea M, Hall JMM, Diakiw SM, Johnston A, Nguyen T, Perugini D, et al. Development of an artificial intelligence-based assessment model for prediction of embryo viability using static images captured by optical light microscopy during IVF. *Hum Reprod* [Internet]. 2020;35(4):770–84. <https://academic.oup.com/humrep/article/35/4/770/5815143>
4. Tran D, Cooke S, Illingworth PJ, Gardner DK. Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer. *Hum Reprod* [Internet]. 2019;34(6):1011–8. <https://academic.oup.com/humrep/article/34/6/1011/5491340>

5. Blank C, Wildeboer RR, DeCroo I, Tilleman K, Weyers B, de Sutter P, et al. Prediction of implantation after blastocyst transfer in in vitro fertilization: a machine-learning perspective. *Fertil Steril* [Internet]. 2019;111(2):318–26. <https://doi.org/10.1016/j.fertnstert.2018.10.030>.
6. Chen T-J, Zheng W, Liu C-H, Huang I, Lai H, Liu M. Using deep learning with large dataset of microscope images to develop an automated embryo grading system. *Fertil Reprod* [Internet]. 2019;01(01):51–6. <https://www.worldscientific.com/doi/abs/10.1142/S2661318219500051>
7. El-Shafeiy E, El-Desouky A, El-Ghamrawy S. An optimized artificial neural network approach based on sperm whale optimization algorithm for predicting fertility quality – studies in informatics and control – ICI Bucharest. *Stud Inform Control*. 2018;27(3):349–58.
8. Chavez-Badiola A, Flores-Saiffe Farias A, Mendizabal-Ruiz G, Garcia-Sanchez R, Drakeley AJ, Garcia-Sandoval JP. Predicting pregnancy test results after embryo transfer by image feature extraction and analysis using machine learning. *Sci Rep* [Internet]. 2020;10(1):4394. <http://www.ncbi.nlm.nih.gov/pubmed/32157183>
9. Akinsal EC, Haznedar B, Baydilli N, Kalinli A, Ozturk A, Ekmekcioğlu O. Artificial neural network for the prediction of chromosomal abnormalities in azoospermic males. *Urol J*. 2018;15(3):122–5.
10. Fauser BC. Towards the global coverage of a unified registry of IVF outcomes. *Reprod Biomed Online*. 2019;38(2):133–7. <https://doi.org/10.1016/j.rbmo.2018.12.001>.
11. Balayla J, Shrem G. Use of artificial intelligence (AI) in the interpretation of intrapartum fetal heart rate (FHR) tracings: a systematic review and meta-analysis. *Arch Gynecol Obstet*. 2019;300:7–14.
12. Boland MR, Polubriaginof F, Tatonetti NP. Development of a machine learning algorithm to classify drugs of unknown fetal effect. *Sci Rep*. 2017;7(1):12839.
13. Bottomley C, Van Belle V, Kirk E, Van Huffel S, Timmerman D, Bourne T. Accurate prediction of pregnancy viability by means of a simple scoring system. *Hum Reprod*. 2013;28(1):68–76.
14. Dyer S, Chambers GM, de Mouzon J, Nygren KG, Zegers-Hochschild F, Mansour R, Ishihara O, Banker M, Adamson GD. International Committee for Monitoring Assisted Reproductive Technologies world report: Assisted Reproductive Technology 2008, 2009 and 2010. *Hum Reprod*. 2016;31(7):1588–609. <https://doi.org/10.1093/humrep/dew082>. Epub 2016 May 20. PMID: 27207175.
15. Gil D, Girela JL, De Juan J, Gomez-Torres MJ, Johnsson M. Predicting seminal quality with artificial intelligence methods. *Expert Syst Appl*. 2012;39(16):12564–73.
16. Iftikhar PM, Kuijpers MV, Khayyat A, et al. Artificial intelligence: a new paradigm in obstetrics and gynecology research and clinical practice. *Cureus*. 2020;12:e7124.
17. Letterie GS, Mac Donald AW. A computer decision support system for day to day management of ovarian stimulation during in vitro fertilization. *Fertil Steril*. 2020;114:1026–31.
18. Liu L, Jiao Y, Li X, Ouyang Y, Shi D. Machine learning algorithms to predict early pregnancy loss after in vitro fertilization-embryo transfer with fetal heart rate as a strong predictor. *Comput Methods Prog Biomed*. 2020;196:105624. <https://doi.org/10.1016/j.cmpb.2020.105624>. Epub 2020 Jun 25. PMID: 32623348.
19. Ma Y, Chen B, Wang H, Hu K, Huang Y. Prediction of sperm retrieval in men with non-obstructive azoospermia using artificial neural networks: leptin is a good assistant diagnostic marker. *Hum Reprod*. 2011;26(2):294–8.
20. Miyagi Y, Habara T, Hirata R, Hayashi N. Feasibility of artificial intelligence for predicting live birth without aneuploidy from a blastocyst image. *Reprod Med Biol* [Internet]. 2019;18(2):204–11. <http://www.ncbi.nlm.nih.gov/pubmed/30996684>
21. Mora-Sánchez A, Aguilar-Salvador DI, Nowak I. Towards a gamete matching platform: using immunogenetics and artificial intelligence to predict recurrent miscarriage. *NPJ Digit Med*. 2019;2:12.
22. Moreira MWL, Rodrigues JJPC, Kumar N, Al-Muhtadi J, Korotaev V. Evolutionary radial basis function network for gestational diabetes data analytics. *J Comput Sci*. 2018a;27:410–7.
23. Moreira MWL, Rodrigues JJPC, Al-Muhtadi J, Korotaev VV, de Albuquerque VHC. Neuro-fuzzy model for HELLP syndrome prediction in mobile cloud computing environments. *Concurr Comput*. 2018b; <https://doi.org/10.1002/cpe.4651>.
24. Nisal A, Diwekar U, Bhalerao V. Personalized medicine for in vitro fertilization procedure using modeling and optimal control. *J Theor Biol*. 2020;487:110105. <https://doi.org/10.1016/j.jtbi.2019.110105>.
25. Paydar K, Niakan Kalhori SR, Akbarian M, Sheikhtaheri A. A clinical decision support system for prediction of pregnancy outcome in pregnant women with systemic lupus erythematosus. *Int J Med Inform*. 2017;97:239–46.
26. Polak S, Mendyk A. Artificial intelligence technology as a tool for initial GDM screening. *Expert Syst Appl*. 2004;26(4):455–60.
27. Tejera E, Joseareias M, Rodrigues A, Ramõa A, Manuelnieto-villar J, Rebelo I. Artificial neural network for normal, hypertensive, and preeclamptic pregnancy classification using maternal heart rate variability indexes. *J Matern Fetal Neonatal Med*. 2011;24(9):1147–51.
28. Thirumalaraju P, Bormann CL, Kanakasabapathy M, Doshi F, Souter I, Dimitriadis I, et al. Automated sperm morphology testing using artificial intelligence. *Fertil Steril*. 2018;110(4):e432.
29. Vickram AS, Kamini AR, Das R, Pathy MR, Parameswari R, Archana K, et al. Validation of artificial neural network models for predicting biochemical

- markers associated with male infertility. *Syst Biol Reprod Med.* 2016;62(4):258–65.
30. Goodson SG, White S, Stevans AM, Bhat S, Kao CY, Jaworski S, Marlowe TR, Kohlmeier M, McMillan L, Zeisel SH, et al 2017. CASanova: a multiclass support vector machine model for the classification of human sperm motility patterns. *Biology of Reproduction* 698–708. (10.1093/biolre/iox120)
31. Mirsky SK, Barnea I, Levi M, Greenspan H, Shaked NT. 2017. Automated analysis of individual sperm cells using stain-free interferometric phase microscopy and machine learning. *Cytometry A* 893–900. (10.1002/cyto.a.23189)
32. Girela JL, Gil D, Johnsson M, Gomez-Torres MJ, De Juan J. 2013. Semen parameters can be predicted from environmental factors and lifestyle using artificial intelligence methods. *Biology of Reproduction* 99 (10.1093/biolreprod.112.104653)
33. Chavez-Badiola A, Flores-Saiffe-Farías A, Mendizabal-Ruiz G, Drakeley AJ, Cohen J. Embryo Ranking Intelligent Classification Algorithm (ERICA): artificial intelligence clinical assistant predicting embryo ploidy and implantation. *Reprod Biomed Online.* 2020 Oct;41(4):585–593. <https://doi.org/10.1016/j.rbmo.2020.07.003>. Epub 2020 Jul 5. PMID: 32843306.
34. Liao S, Pan W, Dai W, et al. Development of a Dynamic Diagnosis Grading System for Infertility Using Machine Learning. *JAMA Netw Open.* 2020;3(11):e2023654. <https://doi.org/10.1001/jamanetworkopen.2020.23654>.



AIM and Gender Aspects in Reproductive Medicine

73

Kiran Heer Kaur

Contents

Introduction	1018
The History of Reproductive Medicine	1018
The Relationship Between AI and Reproductive Medicine	1020
Gender and Reproductive Medicine: Objectives, Issues, and Challenges	1021
Reproductive Medicine and Gender: A Century-First Century Approach	1023
The Male Experience and Reproductive Medicine	1023
Transgender, Gender Fluid, and Nonbinary Perspectives and Reproductive Medicine	1024
Concluding Thoughts	1026
References	1026

Abstract

Artificial Intelligence invites boundless opportunities and cutting-edge innovation in Reproductive Medicine (RM). In the future, machine learning (ML) algorithms will offer effective and efficient utility in assisted reproductive technologies (ART), as well increase the potential for individualized patient care. As the future of RM offers great advancement it is imperative to understand how this will inevitably impact people. Therefore, the purpose of this chapter is to highlight the diverse gender perspectives and issues that could derive from advanced RM. What this entails is observing

how increased technology can impact the future of gender, as ART and future enhancement technologies allow prospective parents the chance to select the gender of their future offspring. It must be examined how these technologies can contribute to existing inequalities, and whether this would worsen gender discrimination and bias. In addition, this chapter aims to employ a twenty-first-century approach to understanding gender aspects in RM. What this involves is highlighting the perspectives of often marginalized groups in RM. This includes highlighting male, transgender, and nonbinary experiences in RM. The point of this is to further enrich the discussion surrounding gender and RM, and make it as inclusive as possible for all people.

K. Heer Kaur (✉)
Center for Social Development, Wolverhampton, UK

Keywords

Artificial Intelligence · Machine learning · Natural language processing · Liberal eugenics · Genetic engineering · CRISPR/Cas9 · Big data · Data security and cybersecurity

Introduction

Artificial Intelligence (AI) is “the ability of machines to learn and display intelligence which is in stark contrast to the natural intelligence demonstrated by humans and animals” [1]. Ever since the invention of Artificial Intelligence and Future Emerging Technologies (FET), pivotal technological breakthroughs have been made all throughout the twentieth and twenty-first centuries. Consequently, the invention of technology has dominated industries across all sectors. As a result, it is imperative to understand how such advances have changed the landscape of modern medicine, with a particular emphasis on reproductive medicine. Thus, the purpose of this chapter is to explore AI, and its applications in medicine, with a thorough and specialized focus on reproductive medicine (RM) and gender perspectives. What this entails is exploring the history and origins of reproductive medicine, understanding the relationship between modern-day AI and reproductive medicine, and the crucial role that gender plays in reproductive medicine. This will be done by exploring how RM has changed and how it will subsequently impact men and women.

As RM will be discussed at length in this chapter, it is essential to establish a definition.

Thus, for all intents and purposes, the definition of reproductive medicine strongly refers that...

Reproductive medicine is a subfield of medicine that specializes in the health of the reproductive organs. This includes the diagnosis, treatment and prevention of infertility in men and women, as well as issues relating to contraception, puberty, menopause and some sexual health related conditions. National Cancer Institute in partnership with the National Institute of Health [2]

As well as establishing a definition for RM, it is equally important to establish the meaning of gender. The meaning of gender in this chapter will explicitly refer to the socially constructed norms, behaviors, and assigned roles associated with masculinity and those that identify as male, and with femininity and those that identify as females [3]. This is not to be confused with biological sex that refers to the biological reproductive organ that one is born with, and assigned at birth [3]. It is also acknowledged in this chapter that there are more than two genders, and thus aims to highlight the experiences of not just biologically born men and women, but to also include the reproductive experiences of transgender and nonbinary people.

The History of Reproductive Medicine

The origins of reproduction theory and ideology can be traced back to Aristotle in ‘*De Generatione Animalium*,’ or as it is more commonly known in the West, *On Generation of Animals*. [4] In this piece of work, Aristotle proposed that human beings are made through the merger of male and female seeds [5]. It was believed that when male sperm came into contact with something with a hematogenous origin, this would cause the development of an embryo from menstrual blood that was present in the female uterus [5]. While there are understandably flaws in this hypothesis, the work of Aristotle is considered to be one of the most important pieces of work and theories in the history of reproductive medicine, and has survived for 2000 years with only slight modifications in its time [5]. In addition to Aristotle, there have been some other key figures that have contributed to the framework and body of literature surrounding reproduction. This can be evidenced by Galen’s conception theory, in which he stated that “seeds come from the fluids of the body...the seed is said to be secreted from each of the parts.” [6] This was of great importance at the time as this theory was a step closer in breaking down the theory of reproduction. Other notable work derives from William Harvey in a book entitled, *Exercitationes de Generatione Animalium* [7]. In

this book, Harvey coined the phrase “ex ovo omnia” which translates to “everything from an egg” [5].

“Ex ovo omnia” [5] was Harvey’s belief that animals are born from a “homogenous mass which has [animal] life actually or potentially” [8], and thus that is how life is formed. Harvey’s theory differed from previous theories as he rejected the hypothesis of Aristotle and Galenists [7], and instead believed that when seminal fluid comes into contact with the female momentarily, it establishes the “fecundity” of the female organism [8]. After some time, the fertilized female forms the initial embryo and then the embryo continues to grow and develop into a human being [8].

These early ideologies of reproductive medicine, while flawed or slightly inaccurate in nature, were very important in laying out the foundation for our current understanding of reproduction, and subsequently the specialist medical field of reproductive medicine.

As scientists continued to theorize about the act of reproduction and fertilization, with it came a curiosity to see if life could be made from outside an organism and with some human intervention. The earliest experiments of in vitro fertilization (IVF), an experimental procedure to see if fertilization could take place outside the body, could be traced back to the 1770s, as Lazaro Spallanzani discovered that frog oocytes developed into tadpoles after having been in contact with semen [5]. In 1827, Karl Ernst Von Baer discovered that mammals also possess oocytes and published his research [5]. Major strides continued to be made and by 1878 research conducted by, Samuel Leopold Schenk, confirmed that cell division could be made outside of the body [5] after IVF experiments were conducted on mammals, such as rabbits and guinea pigs.

All of these experiments and findings were of great importance and served as the foundations for doctors and pioneers, such as John Rock, Miriam Friedman Menkin, and Min Chueh Chang, to start conducting experiments in human beings [5]. After extensively researching fertilization outside the body in mammals, the next goal and breakthrough for the scientific community was to

see if fertilization could take place in humans outside the body. This breakthrough was achieved in 1944, when Dr. Rock and Miriam Menkin found that a human egg could be fertilized outside the womb [9]. They then published their findings in an edition of *Science Magazine* entitled “In Vitro Fertilization and Cleavage of Human Ovarian Eggs” [9]. At the time of the publication the news had become somewhat of a contentious topic and had garnered significant attention from scientific, public, and religious circles.

Noteworthy to mention is that one of the most interesting perspectives on this topic derived from couples, particularly women, who had struggled with infertility. Numerous women had written directly to Dr. Rock after the publication and had inquired about what the procedure was, what it entailed, and if it “could help them to become pregnant?” [10] A reporter, Joan Younger, who had been following the research of Rock and Menkin had written an article, “*Life begins in the test-tube*” [11], in which she had discussed an increase in infertility among healthy couples, and how this new procedure was a symbol of hope. Women at the time were happy to volunteer themselves as test subjects for Rock’s experiment [11]. This demonstrates the clear attitudes women had very early on about IVF and how for many it was seen as a renewal of hope. Although this news had encouraged hope for many infertile couples, the experimental procedure received criticism from religious circles. In many religions, the idea of creating life outside the human body was considered to be “unnatural,” and going against the Will of God [12].

In the twenty-first century, the introduction of Artificial Intelligence (AI) and future emerging technologies (FET) in medicine has raised similar concerns and sentiments that people have had in the past about IVF, and Assisted Reproductive Technologies (ART). This is because some people are skeptical about the use of AI and FET in medicine, with regard to reproductive medicine in particular. Contention on this topic derives from controversial debates surrounding ideas around genetic engineering/modification and biotechnology. With technology being a catalyst to make these procedures and innovations possible, some

people fear what this could mean for future generations. This very topic has since accumulated much literature on both sides. To adequately understand the different perspectives on AI in reproductive medicine, it is important to ask two key questions:

1. What is the relationship between AI and Reproductive Medicine?
2. What are the future objectives, issues, and challenges that derive from this?

These aforementioned questions will be discussed at length.

The Relationship Between AI and Reproductive Medicine

Over the past few years significant advancements in AI have led to the widespread implementation of AI in medicine [13]. AI applications such as ML, NLP, and robotic technology have helped to revolutionize the field of medicine by assisting doctors in the prognosis, diagnosis, and treatment of their patients, and have largely improved patient care and satisfaction. Such advancements and improvements in medicine have been possible through ML algorithms that allow computers to learn, compute, and analyze trends and patterns based upon large and complex datasets [13].

ML techniques and applications have proven to be vital in health care as sufficient evidence from science-based studies present an increase in accuracy of these applications, aiding doctors and their prognosis of patients. This phenomenon can be evidenced by a unique case study, in which a newborn baby that appeared to be healthy continued to have constant seizures, most commonly referred to as status epilepticus; despite having no signs of infections and clean CT scans [14]. The prognosis for this newborn baby appeared bleak, as numerous drugs to reduce and control the seizures had failed, and the seizures had become even more pronounced [14]. Eventually a blood sample of the infant was sent for rapid whole-genome sequencing, at Rady's Genomic Institute [14]. "The sequence encompassed 125 gigabytes of data" [14] and had searched

almost "five million locations where the infant's genome differed from the most common one" [14]. Within 20 s the NLP AI application had processed the infants digital medical record and had determined "eighty-eight phenotype features" [14] (observable traits and characteristics of an organism). The ML algorithms were utilised to filter through five million genetic variations to find the rarest gene variations that could cause the onset of these seizures [14]. Using the child's phenotypic data and cross-referencing it with all the possible genetic variations, the system was able to identify the gene ALDH7A1, as the most likely cause for the seizures [14]. The gene variant is extremely rare and "occurs in less than 0.01% of the population" [14]. After the detection of this extremely rare variant, which due to its rarity could have been overlooked by doctors, the infant was able to seek treatment and was healthy overall [14]. This case highlights firsthand how impactful and beneficial AI can be in healthcare.

In the field of RM, AI is proving to be extremely helpful in enhancing ART and can actually serve as a solution to overcome some of the biggest challenges pertaining to infertility in couples [1]. As AI and ML algorithms have the potential to analyze large and complex datasets, as previously described in the aforementioned case study, this could significantly change and impact the field of RM. With the ability to analyze large datasets in real-time, the use of AI applications could result in more personalized treatment options for fertility patients [15]. As various patient information such as different fertility issues, solutions, treatment plans, procedures, and outcomes can be analyzed and cross-referenced in databases, this can effectively assist doctors. In fact, this has huge significance as not only can diagnoses be made faster but the appropriate and individualistic treatment options can then be given to patients [15].

In a 2019 article, *Artificial Intelligence: its applications in reproductive medicine and the assisted reproductive technologies*, the collective authors discuss AI and ART, and hypothesize that the "application of AI in IVF will be significant" [15]. This is because AI can be used in the "assessment of ovarian reserve parameters and sperm selection" [15], and in the evaluation and selection

of embryos [15]. The fact that AI can be utilized to detect pivotal indicators and “developmental hallmarks” [15] for viable embryos is especially of significant importance as it allows for greater accuracy and hence, increases the risk of successful insemination [15]. The use of AI promises greater accuracy and results in embryo selection as current methods rely on the analysis of embryologists reviewing a combination of factors [15]. The issue with this methodology, however, is that there is no standardized approach in the analysis of viable embryos and analysis can thus be largely subjective [15].

This very phenomenon became apparent after a study highlighted that different embryologists had varying results in embryo grading and viability; thus, presenting how embryo selection was highly subjective [15]. Breakthroughs in AI and ML programs have allowed for computers to input the data of embryos and determine a standardized approach to grade the viability of embryos [16]. The KIDSscoreD5 system is an AI-powered system that accurately studies, analyzes, assesses, and classifies embryos [16]. Thus, this system can increase the likelihood of successful implementation and birth. A series of studies has further shown that the KIDSscoreD5 system has had a significant impact in ART and has shown great clinical promise [16]. The KIDSscoreD5 system was able to effectively analyze “more than 20,000 embryos in more than 3000 patients” [16]. Not only was the system able to compute the data of the embryos but was able to do so with effective results, thus increasing the possibility of selecting the best embryo for fertilization [16]. Moreover, the KIDSscoreD5 system is able to select the embryo with the least likely chance of having a chromosomal mutation [16]. The gains made in ART as a result of advanced AI have been immense and continues to change and innovate remedial tasks, thus, resulting in better and more favorable outcomes for patients.

Gender and Reproductive Medicine: Objectives, Issues, and Challenges

As reproductive medicine continues to accelerate and develop at a rapid pace due to the prevalence

of technology, the future of reproductive medicine is full of possibilities and challenges. The way reproductive medicine will change and develop in the future will significantly impact the population and has the potential to be a double-edged sword for modernity and freedom of choice. This is largely because reproductive medicine can often be associated with choice, and the freedom to have choices, when it comes to one’s reproductive organs and fertility. However, as technology allows us to potentially expand our choices, it must be imperative that people make safe and responsible choices when using this technology.

As AI and technology becomes increasingly more embedded within the field of RM, this will inevitably change the nature of RM. As technology becomes smarter, this opens the door to a number of possibilities in this domain. One of the areas that will greatly and fundamentally benefit will be genetic medicine. As it is desired that in the future, genes can be edited or altered via genetic engineering. “Genetic engineering (GE) is a process that involves changing, altering or manipulating an organism” [17]. GE is a broad term and also encompasses many synonyms such as genetic modification, genetic manipulation, genetic technology, biotechnology, and recombinant DNA technology [17]. There are many purposes and practical functionalities of GE [17]. This includes isolating and researching gene properties, structures, and functions [17], producing “useful proteins by novel methods” [17], diagnosing and treating patients [17], and “genome analysis by DNA sequencing.” [17] In order to do this, a gene editing technique, clustered regularly interspaced short palindromic repeat (CRISPR)-associated system (Cas) technology [18] is employed. CRISPR allows genetic engineers and doctors to precisely select, “target and study certain DNA sequences in the vast expanse of the genome.” [19] CRISPR-Cas9 also allows scientists the possibility to edit DNA [19], as through this technology, it is possible to manipulate genes and send “proteins to specific DNA targets” [19] that can switch genes on or off, or even completely “engineer entire biological circuits” [19]. The way that CRISPR-Cas9 works to make gene editing possible is through the protein Cas9. Cas9 functions as “molecular

scissors” [20] and can be programmed to cut a specific part of the DNA “as it carries with it Guide RNA (gRNA)” [20]. The gRNA will guide Cas9 exactly where to cut the DNA from [20]. After the cut has been made in the DNA, the DNA will try to rebuild itself; however, this is where scientists can manipulate and alter this process [20]. Scientists can instead introduce “template DNA that guides what bases are inserted in the gap.” [20] This then results in the genes being edited. GE using CRISPR-Cas9 thus far has mostly been successful in plants and animals [21]. In humans beings, GE has been largely experimental and is often met with contentious moral, ethical, and philosophical debates.

With the potential to edit genes, this raises some serious questions about the direction we are heading in and what it is that we want to achieve in the future with such technology. Many people believe that the ultimate goal will be the “designer baby.” [22] A designer baby is a baby that has been genetically enhanced by the pre-selection of genes prior to birth. What this connotes is that, genes that are inherently linked to more desirable superficial characteristics and traits, for example, blue eyes, brunette hair, and height, can be selected to ultimately create a genetically enhanced version of a child [22].

Some fertility clinics across the USA offer the opportunity to prospective parents to select certain attributes for their offspring through preimplantation genetic diagnosis (PGD) [23]. PGD sometimes referred to as PGT-M is a treatment option that allows embryos to be screened for genetic/chromosomal abnormalities [24]. In addition to this, PGD can also reveal other important genetic information regarding the embryo such as, the likelihood of the embryo being male or female, or even the child’s eye color. The Fertility Institutes is a fertility practice that offers the opportunity of gender selection to prospective parents [24]. The way that the procedure is done is firstly by extracting the eggs from the mother and collecting the sperm from the father [23]. The eggs are then fertilized via IVF in a laboratory [23]. The fertilized eggs are then tested and screened for genetic or chromosomal irregularities and also the desired gender [23]. A genetic test will reveal the likelihood of having a baby of a

particular gender or feature for example, eye color [25]. The selected embryo will then be implanted into the mother and gestation and birth will proceed as normal [23].

If in the future parents have permission and accessibility to utilize this technology, it leads to some wider concerns about the power and choices parents could make regarding their child. This is because this technology would give parents the freedom to choose certain attributes of their future offspring, including, even, the gender of their child. This may have some potentially adverse consequences for people that favor a particular gender over another. While the gender of a child may seem trivial at face-value, this issue is a deep-seated one that remains in the conscious bias of many cultures, globally. Particularly, many Asian cultures tend to favor males over females [26]. In Asian countries, such as China and India, patriarchal norms and ideologies are widely imbedded within society with there being an increased value placed both socially and economically on males [26]. Consequently, women and girls endure severe mistreatment both by society and familial structures [26]. Furthermore, numerous studies highlight a new and alarming trend in which baby girls in India and China are becoming victims of gendercide [26].

Gendercide is “the systemic annihilation of female fetuses solely for their sex in preference of their male counterpart.” [26] This systemic discrimination and bias has resulted in two hundred million girls less in the world, as these baby girls were “killed, aborted and abandoned” [26] solely because of their gender [26]. The ratio of girls to boys in these countries continue to decrease and this has been a constant trend since the 1990s [26]. Despite these alarming concerns, the Indian and Chinese governments have not taken measures to remedy this problem; nor has there been sufficient efforts to adequately address and or publicize the issue [26]. This apparent discrimination and declining gender issue are important to consider at present, as it must be considered what could and would happen, if parents had the freedom to select the features for their potential offspring. In countries like China and India, where gendercide is already a dangerous and concerning problem, it must be considered

what would happen if genetic editing technology was incorporated into the mix. This aspect raises many questions such as:

- Would the availability of gene editing technology be knowingly contributing to already established social problems and gender-based discrimination?
- How can it be ensured that state governments can regulate the responsible and safe use of this technology, when countries like China and India have already been unresponsive to the gendercide that has been happening in their respective countries?
- What rights should the potential offspring have? Is it violating the rights of the child to choose their gender as a novel reason; for example, financial gain, social acceptance, and carrying family names?

Reproductive Medicine and Gender: A Century-First Century Approach

Traditionally, RM was commonly associated with the experience of those born with female reproductive organs. While infertility affects both men and women, and the procedures that derived from modern medicine have benefited men and women, the issues surrounding biologically born male RM are still lacking in the literature [27]. Despite the large role that men play in reproduction, the male experience within RM is mostly absent. However it is not just biologically born men whose experiences are lacking, as transgender and nonbinary perspectives are also excluded. As our knowledge of gender in the twenty-first century evolves, it is important to include all the diverse gender perspectives. Thus, the purpose of this section will be to adequately discuss reproductive medicine from the male experience and from transgender and nonbinary perspectives, respectively.

The Male Experience and Reproductive Medicine

RM as a discipline has largely been synonymous with women and, thus, has largely excluded men

from this narrative. This is because within the field of RM, fertility is one of the leading concerns and areas of treatment. However throughout history, fertility has been mainly associated with women. This can be demonstrated through the depiction of women in religion, and throughout Greek and Egyptian mythology and folklore. The religious link between women and fertility can be presented through Eve, a figure first depicted in the Book of Genesis [28]. Eve is most commonly associated with the forbidden apple that she had taken from the Garden of Eden, against Gods wishes; resulting in God punishing Eve with agonizing childbearing and labor [28]. In the Hindu religion, fertility is associated with Parvati, the ultimate divine mother and feminine symbol [29]. The link between fertility and women is not isolated strictly to religion, as Greek and Egyptian mythology also feature goddesses of fertility. This is evident through the Egyptian Goddess, Isis and the Greek Goddess, Aphrodite. The presence of all these figures throughout history has emphasized the relationship between fertility, motherhood, and women.

As the understanding of fertility and reproduction evolves and medical procedures and studies clarify what causes fertility issues in both men and women, this greatly alters the ontological knowledge and understanding on this topic. Men equally play an important part in the study and practice of RM; yet at times their experiences do not receive as much attention as their female counterparts. The use of ART has helped to improve some of the issues involving men's reproductive health and infertility; however, a lot still remains unknown in this domain. While ART procedures such as intrauterine insemination and intracytoplasmic sperm injection have helped to overcome some of the barrier's men face when trying to conceive a child, [30] there is still a significant lack of evidence-based clinical study and knowledge [27]. The lack of knowledge is especially concerning as studies have shown a global trend in "declining sperm counts" [31] and further abnormalities in the male reproductive system [31]. It is further indicated that male infertility ensues in "40% of couples experiencing infertility" [31]. As a result, men's declining reproductive health is considered to be a "global

crisis” [31]. The reason behind this crisis is due to a lack of public awareness concerning men’s reproductive health, a lack of proficient research and funding in this area, the hesitance of men to get examined, and inadequate healthcare policies [31]. The statistics and factors regarding the decline in men’s health is of grave importance as reproductive issues in men has also been found to have a positive correlation to other health problems, including “diabetes mellitus, metabolomic disorders and cardiovascular disorder” [31].

When considering the inequality between male life expectancies in comparison to female life expectancies [31], this link is important to depict. If there were more awareness and encouragement for men to get examined and discuss their health issues, this would be a start in breaking down the crisis. Accompanying this, further research would be needed to understand the pathophysiological basis of male subfertility and infertility [27]. In a study conducted by researchers at Stanford University School of Medicine, it was deemed possible that during fetal development any harmful environmental factors could contribute to reproduction problems, as well as other health concerns [32]. However, further evidence and examination would be needed to identify the links between reproductive issues and overall health in men. The emergence of AI in RM could bring fundamental benefits to men and male health, however before technology can improve the difficulties pertaining to male health, the issues that cause this need to be further explored.

Transgender, Gender Fluid, and Nonbinary Perspectives and Reproductive Medicine

Although transgender, gender fluid, and nonbinary medicine can be considered to be its own emerging interdisciplinary area of study and medical specialty, when discussing gender perspectives and RM it would be appropriate to recall that gender conceptually encompasses more than just traditional male and female perspectives. The meaning of gender in the twenty-first century is an essentially contested topic. This is because

traditionally gender would be synonymous with cisgender men and women. Cisgender refers to a person whose “gender identity is the same as the one assigned at birth.” [33] However, the coming out of celebrities such as, Caitlyn Jenner, as transgender male to female and singer songwriter, Sam Smith, coming out as nonbinary has set a new precedent for gender and the public understanding of it over the past few years [27].

As our collective knowledge and consciousness evolves around the areas of gender and biologically assigned sex, it is important to understand that the two are not always mutually exclusive. Therefore, in order to discuss the various gender perspectives in regard to RM, it would be insightful to include perspectives from all these underrepresented groups.

Currently literature surrounding this topic is thin; however, in a 2020 article entitled “*The imperative for Transgender and Gender non-binary inclusion*” [33], the authors made a strong case that by solely referring to the experiences of cisgender women in reproductive health, this excludes and marginalizes an entire group of people that have reproductive health and medical concerns that are important and unique to share [33]. This viewpoint is an accurate depiction of the current literature as transgender and nonbinary perspectives are largely excluded from this topic.

The perspectives of transgender and nonbinary people in relation to AIM and reproductive medicine is important for three reasons that will be discussed in this section. Firstly, this gender aspect is important to include as by not doing so would be short-sighted and exclusive to entire groups of people that make up an important proponent of society. To be exclusive to these various gender aspects would also “prevent the advancement of science and clinical care for people of all genders.” [33] Secondly, it is important to include transgender and nonbinary perspectives in order to mitigate the possibility of technological bias in the future. Thirdly, it is important as with the breakthroughs that are being made in RM, as a result of AI and other emerging technologies, this can directly impact the futures for many transgender and nonbinary individuals who may use ART to have children.

As previously mentioned when discussing the lack of inclusivity of male reproductive issues, the perspectives of transgender and nonbinary people are also excluded in the mainstream literature as traditionally issues surrounding RM and infertility were often associated with cisgender women [33]. This is important as the lack of adequate research and knowledge on other genders and gender identities could prevent doctors from making more informed diagnoses in regard to health care. For example, without sufficient research and acknowledgment of these groups, it would make it more difficult to establish links between certain illnesses and diseases, thus patients may not receive the right or best care that they possibly could.

According to a report by the World Health Organization (WHO), gender plays a pivotal role in “people’s experience of and access to healthcare.” [3] This is because gender can either limit or improve one’s access to healthcare services, support, resources, and information [3]. Literature on transgender and nonbinary health care suggested that transgender and nonbinary people are less likely to request healthcare services citing discrimination, inequality, and an overall lower quality of health care in comparison to their cisgender counterparts as the reason [33]. Harmful gender bias can have a profound impact for transgender, nonbinary, and non-gender conforming groups, as the lack of empirical evidence in understanding health problems that these individuals could be susceptible to having would continually lead to misdiagnosis’ and poor access to health care. Unless more research, studies, and conversations take place to make these groups visible in all aspects of medicine, inequality will continue to persist. This underlying bias and discriminatory behavior will also hinder the progression and improvement of overall health care including RM.

Leading on from this, it is also important that these gender issues are taken into consideration as the implementation of technology can worsen existing bias. For example, as health care in general continues to advance through the innovation of AI and ML tools, it is important we do not have medical apps or programs that purposely exclude transgender and nonbinary people. The

implementation of AI and ML applications may lead to more unintended bias. This could be possible as there is evidence indicating a lack of empirical research and data in transgender and nonbinary medicine [34]. People with diverse gender identities are said to be “underrepresented in medical research” [34] due to bias, discrimination, and stigma around gender identity [34]. As previously mentioned, the discrimination that transgender and nonbinary people face leads to a harmful cycle in which patients feel uncomfortable seeking treatment, thus, less is known about certain medical issues, illnesses, and diseases that these people could face. This very issue is raising concerns that the absence of sufficient data used for “ML algorithms may contribute to socioeconomic disparities in healthcare” [35].

In order to reconcile the lack of data from those with various gender identities, researchers at Stanford University are running *The Population Research in Identities and Disparities for Equality* (PRIDE) study [34]. The PRIDE study created a digital platform for those who identify as a “sexual or gender minority” [34] with the aim to improve representation from underrepresented groups; and to collate large-scale data sets [34]. Studies such as the one conducted by PRIDE are especially helpful in breaking down barriers and stigmas surrounding gender. As health care gets increasingly more digitalized, it is crucial to make it as inclusive as possible for all people.

Furthermore, transgender and nonbinary perspectives are incredibly important and relevant in the field of RM. ART procedures have made it possible to reinvent parenthood so that LGBTQ and non-gender conforming individuals can have children. Fertility preservation of transgender people is becoming increasingly more common as procedures such as egg freezing, sperm freezing, embryo freezing, and ovarian tissue cryopreservation allow transgender people the opportunity to have biological children in the future [36]. When thinking about the future of RM and the achievements that can be made through AI, the successes will be beneficial to all people in society that would want to, and or choose to, have children. As AI has ample potential to transform RM,

it is also important to consider the issues discussed earlier within this section, regarding invisibility in health care and potential data bias. By ensuring that there are substantial discussions and data, it should be of equal importance that transgender and nonbinary people also have accessibility to ART and other reproductive procedures and technologies.

Concluding Thoughts

AI will undoubtedly play a large role in the future of RM. The technological developments that are being made have the potential to improve and remedy some of the most pressing issues in RM, especially pertaining to fertility. The use of AI and ML to aid ART procedures offers a realm of possibilities for prospective parents. Much like the early conceptualized views people had about RM, and procedures such as IVF, this brings with it a renewal of hope for patients. The implementation of AI can help to treat patients' health concerns via a more individualistic and scientific approach.

While this invites innovation, it is important to remember the full impact that AI can have on the future of RM; especially with regard to gender. It is important that these technologies are not misused to push further gender inequality through the intentional selection of gender in embryos. It should be remembered that while technology can help improve the quality of life for people and help them in their fertility journey, the aim is not to worsen existing social issues such as the current gendercide issue.

In addition, when thinking about gender and RM, it is important that the definition of gender evolves to encapsulate all genders and all people. In order to ensure RM truly advances, it is not technology alone that will achieve this endeavor but our collective thinking on the subject. What this entails is adopting a truly twenty-first-century approach to viewing gender and RM. This includes expanding and widening our definition of gender to include all diverse perspectives and understanding that fertility issues are not synonymous with just the female experience. Much like

AI, gender will also play a large role in the future of RM. Thus, it is important to ensure that gender is considered in every process of RM, including the perspectives from those that identify as a different gender than the traditional ones.

References

- Wang R, et al. Artificial Intelligence in reproductive medicine. Society for Endocrinology. 2019. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6733338/>. Accessed 15 Nov 2020.
- National Cancer Institute. Reproductive medicine. National Institute of Health; 2020. <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/reproductive-medicine>. Accessed 17 Nov 2020.
- World Health Organization [WHO]. Gender and health. https://www.who.int/health-topics/gender#tab=tab_1. Accessed 10 Nov.
- Kremer J. The Haematogenous reproduction theory of Aristotle. Researchgate Nederlands tijdschrift voor geneeskunde. 2004;147(51):2529. <https://pubmed.ncbi.nlm.nih.gov/14735853/>
- Empowered Women's Health. The history of IVF: from folklore to technological marvel. <https://www.volusonclub.net/empowered-womens-health/the-history-of-ivf-from-folklore-to-technological-marvel/>. Accessed 18 Nov 2020.
- Boylan M. Galen's conception theory. *J Hist Biol*. 1986;19(1):47–77.
- Harvey W. *Exercitationes de Generatione Animalium*. London: typis Du-Gardianis; 1651.
- Foote ET. Harvey: spontaneous generation and the egg. *Ann Sci*. 2006;25(2):139–64. <https://doi.org/10.1080/0033796900200071>.
- Rock J, Menkin M. In vitro fertilisation and cleavage of human ovarian eggs. *Science Magazine*, V.100 I.2588 August 1944, p. 105–7.
- Davis G, Loughran T. *The Palgrave handbook of infertility in history: approaches, contexts and perspectives*. Basingstoke: Palgrave Macmillan; 2017.
- Younger J. Life begins in a test tube. *Collier's Weekly*, 10th March 1945.
- Anderson ML, Walker AT. Breaking Evangelicalism's silence on IVF. <https://www.thegospelcoalition.org/article/evangelicalisms-silence-ivf/>. Accessed 18 Nov 2020.
- Wang R, et al. Artificial intelligence in reproductive medicine. Society for Endocrinology. 2019. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6733338/>. Accessed 20 Nov 2020.
- Topol E. Deep medicine: how artificial intelligence can make healthcare human again. New York: Basic Books; 2013.
- Zaninovic N, et al. Artificial intelligence: its applications in reproductive medicine and the assisted

- reproductive technologies. *Fertil Steril.* 2019;112(1):28–30. <https://doi.org/10.1016/j.fertnstert.2019.05.019>.
16. IVIRMA. Artificial Intelligence in embryo selection: a reality thanks to IVIRMA Global. <https://www.ivirmainnovation.com/artificial-intelligence-embryo-selection>. Accessed 30 Nov.
17. Reiss MJ, Straughan R. Improving nature? The science and ethics of genetic engineering. Cambridge: Cambridge University Press; 1996.
18. Liang P, et al. CRISPR/Cas9-mediated gene editing in human triploid zygotes. *Protein Cell.* 2015;6(5):363–72. <https://doi.org/10.1007/s13238-015-0153-5>.
19. Ledford H. CRISPR: gene editing is just the beginning. <https://www.nature.com/news/crispr-gene-editing-is-just-the-beginning-1.19510>. Accessed 23 Dec 2020.
20. NOVA PBS Official. September 9th 2020. The realities of Gene editing With CRISPR | NOVA | PBS. YouTube Video. <https://www.youtube.com/watch?v=E8viPdGrKg>. Accessed 16 Dec 2020.
21. National Research Council (US) Committee on identifying and assessing unintended effects of genetically engineered foods on human health ‘safety of genetically engineered food: approaches to assessing unintended health effects.’ Washington, DC: National Academies Press; 2004.
22. Fukuyama F. Our posthuman future: consequences of the biotechnology revolution. London: Profile Books; 2003.
23. The fertility Institute. Select the gender of your next baby. <https://www.fertility-docs.com/programs-and-services/gender-selection/select-the-gender-of-your-baby-using-pgd.php>. Accessed 28 Nov.
24. Human Fertilisation & Embryology Authority. Pre-implantation genetic diagnosis (PGD). <https://www.hfea.gov.uk/treatments/embryo-testing-and-treatments-for-disease/pre-implantation-genetic-diagnosis-pgd/>. Accessed 28 Nov.
25. Unnatural Selection. [Online] Leeor Kaufman. United States. Radley Studios. Date viewed 26 Dec 2020. Netflix.
26. Ufret S. No one wants a baby girl: analyzing gendercide in China and India. *Glob Majority E-J.* 2014;5(2):117–27. https://www.american.edu/cas/economics/ejournal/upload/ufret_accessible.pdf. Accessed 29 Nov 2020.
27. Bhasin S. A perspective in the evolving landscape in male reproductive medicine. *J Clin Endocrinol Metab.* 2016;101(3):827–36. <https://doi.org/10.1210/jc.2015-3843>.
28. Greneauxgardens. Eve’s curse: a biblical look at Miscarriage and infertility. <https://www.greneauxgardens.com/greneauxgardens/2014/10/eves-curse-biblical-look-at-miscarriage.html>. Accessed 29 Nov 2020.
29. Anon. The Hindu Goddess Parvati. <https://thegoddessgarden.com/the-hindu-goddess-parvati/>. Accessed 29 Nov 2020.
30. Tournaye H. Management of male infertility by assisted reproductive technologies. *Baillieres Best Pract Res Clin Endocrinol Metab.* 2000;14(3):423–35. <https://doi.org/10.1053/beem.2000.0089>.
31. De Jonge C, Barratt CLR. The present crisis in male reproductive health: an urgent need for a political, social and research roadmap. *Andrology.* 2019;7(6). <https://doi.org/10.1111/andr.12673>.
32. Stanford Medicine. Infertile men have a higher risk of heart disease, study finds.’ <https://med.stanford.edu/news/all-news/2015/12/infertile-men-have-a-higher-risk-of-heart-disease-diabetes.html>. Accessed 28 Nov 2020.
33. Moseson H, et al. The imperative for transgender and gender nonbinary inclusion. *Obstet Gynecol.* 2020;135(5):1059–68. <https://doi.org/10.1097/AOG.0000000000003816>.
34. Fierce Healthcare. PRIDE Study uses smartphone apps, digital platform to build cohort for LGBTQ health research. <https://www.fiercehealthcare.com/tech/pride-study-uses-smartphone-apps-digital-platform-to-build-cohort-for-lgbtq-health-research>. Accessed 30 Nov.
35. Gianfrancesco MA, et al. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med.* 2019;178(1):1544–7. <https://doi.org/10.1001/jamainternmed.2018.3763>.
36. Complete Fertility Centre. Transgender fertility and preservation and treatment. <https://www.completefertility.co.uk/fertility-treatments-services/fertility-preservation/transgender-fertility-preservation-and-treatment>. Accessed 4 Jan 2021.



Artificial Intelligence in Pediatrics

74

Christopher J. Kelly, Alexander P. Y. Brown, and James A. Taylor

Contents

Introduction	1030
Challenges in Pediatric AI	1031
Recent Developments in Pediatric AI	1032
Cardiology	1032
Respiratory	1033
Genetics	1035
Endocrinology	1035
Neonatology	1036
Ophthalmology	1038
Primary Care	1039
Radiology	1039
Pediatric Intensive Care	1040
Gastroenterology	1040
Future Potential for AI in Pediatrics	1040
References	1041

Abstract

Pediatrics is a specialty with significant promise for the application of artificial intelligence (AI) technologies, in part due to the richness of its datasets, with relatively more complete longitudinal records and often less heterogeneous patterns of disease compared to adult medicine. Despite considerable overlap with adult medicine, pediatrics presents a distinct set of clinical problems to solve. It is tempting to

assume that AI tools developed for adults will easily translate to the pediatric population, where in reality this is unlikely to be the case. The challenges involved in the development of AI tools for healthcare are unfortunately exacerbated in pediatrics, and the implementation gap between how these systems are developed and the setting in which they will be deployed is a real challenge for the next decade. Robust evaluation through high quality clinical study design and clear reporting standards will be essential. This chapter reviews recent work to develop artificial intelligence solutions in pediatrics, including developments across cardiology,

C. J. Kelly (✉) · A. P. Y. Brown · J. A. Taylor
Google Health, London, UK

respiratory, gastroenterology, neonatology, genetics, endocrinology, ophthalmology, radiology, pediatric intensive care, and radiology specialties. We conclude that AI presents an exciting opportunity to transform aspects of pediatrics at a global scale, democratizing access to subspecialist diagnostic skills, improving quality and efficiency of care, enabling global access to healthcare through sensor-rich Internet-connected mobile devices, and enhancing imaging acquisition to reduce radiation while improving speed and quality. The ultimate challenge will be for pediatricians to find ways to deploy these novel technologies into clinical practice in a way that is safe, effective, and equitable and that ultimately improves outcomes for children.

Introduction

One of the earliest applications of artificial intelligence (AI) in pediatrics was published in 1984, and described a system called SHELP that could diagnose inborn errors of metabolism [1]. Since these early beginnings, recent advances in AI have led to considerable excitement about its potential for improving healthcare [2,3]. Pediatrics and adult medicine share many of the same subspecialties, and consequently there is considerable overlap in terms of potential applications of AI in both. However, pediatrics presents a unique set of clinical problems to solve, and development of AI systems for children involves a range of specific challenges. As a result, in order for children's health to benefit appropriately from these new technologies, a thriving field of pediatric-specific AI research and product development is required.

It is tempting to assume that AI tools developed for adults will easily translate to the pediatric population. In reality, this is unlikely to be the case. Successful applications of AI to clinical problems in pediatrics will require a deep understanding of the challenges and potential pitfalls, and careful external validation in this setting. The importance of this was demonstrated in one study where two vertebral fracture detection systems

designed for adults were applied to pediatric spine X-rays [4]. The overall sensitivity dropped significantly from 98% to 26–36%, and specificity dropped from 99% to 95–98%, rendering these adult systems useless as a screening tool for children.

While recent highest profile developments in AI for health have largely been outside of pediatrics, there have been promising signs of growth in pediatric AI research. A recent systematic review identified 363 studies where a machine learning (ML) model was assessed for the diagnosis, prediction, or management of a condition in children and adolescents [5]. Only a handful of studies used deep learning methods ($n = 15$, or 4%) that have contributed to many significant advances in medical imaging AI over the past decade. Major contributions came from subspecialties including neonatology and neurology, while most studies originated from high-income (82%) and upper middle-income (15%) countries. Only 3% came from lower middle-income countries, with just three studies (0.8%) coming from the African continent.

Pediatrics is arguably one of the most promising specialties for AI researchers to explore. In the intensive care specialties, large amounts of data are generated (estimated approximately 1 terabyte of data per neonatal intensive care bed per year [6]), including data types where AI methods have been shown to be useful, such as imaging, electronic health record (EHR), physiological time series, and genetic data. Children more commonly become unwell due to a single disease process, as compared to the polypathology and accompanying polypharmacy often found in adult populations. Past medical histories are also often naturally simpler than with adults. As a result, the pediatric patient population is generally less heterogeneous, making machine learning projects potentially more straightforward to design and execute.

Pediatric health records are also relatively more complete than those for adults. This is particularly true in the neonatal intensive care unit (NICU), where infants are usually admitted from birth or shortly thereafter, with close monitoring of physiological parameters, feeding, interaction,

test results, medication administration over their entire lives thus far. These high-fidelity datasets are unrivaled within medicine for their completeness and accuracy. Following discharge from NICU, infants are often followed up for several years, resulting in the availability of high quality longitudinal outcomes. Children are also more likely to have significant portions of their health record already digitized, given the increasingly widespread adoption of EHRs over the past decade.

In terms of impact, the benefits of improving health in children translates into some of the highest quality adjusted life years (QALYs) possible across medicine. This is because small improvements in mortality or morbidity have an effect over the rest of a child's life. Development and deployment of AI tools may democratize and expand global access to high-quality diagnostics and medical advice, particularly in low resource settings where age distributions are skewed toward younger age groups. Approximately, 50% of the population is under 18 years of age in much of Subsaharan Africa, along with countries such as Iraq, Yemen, and Afghanistan, compared with 15–25% in developed countries [7]. In combination with explosive growth in internet-enabled mobile devices, which present new ways to reach remote communities with historically poor access to healthcare, pediatric AI may offer promising solutions to a range of important global child health needs.

This chapter aims to provide an overview of recent developments in AI in pediatrics, explore the benefits, and challenges of working with pediatric data, and outline a vision for the future promise of these exciting new technologies to improve child health.

Challenges in Pediatric AI

All research in pediatrics – whether or not AI is involved – encounters the particular challenge of ensuring that the use of data derived from children, who may not be able to meaningfully consent to participate in research, adheres to legal and ethical standards. While legal guardians may

consent on behalf of children, this process has historically required additional oversight and regulation of pediatric data collection and interventional studies. As a result, access to pediatric datasets involve more complex and longer approval timelines than those obtained from adults, with very few publicly available datasets. Furthermore, disease patterns in children are closely coupled to age, meaning that only a portion of a given dataset may be useful for a particular application, further reducing the statistical power available from an already small population.

In the pharmaceutical industry, the reputational and legal ramifications of any error or negative outcome in trials involving children, alongside the additional cost and time required to run pediatric studies, has led historically to many medications only being tested and licensed for adult patients. This changed in the early twenty-first century, with the introduction of US, EU, and Canadian legislation requiring trials for novel and existing medications to consider pediatric applications. Reputational and legal risks in the development of AI parallel those of medication or diagnostics development, and legal protections as well as regulatory incentives may be necessary to ensure that children can also benefit from advances in AI.

Even if the development of AI tools in a pediatric context is more challenging, other benefits could tip the balance in favor of this work. Where disease can be prevented or effectively treated in a child, the health economic implications of this in terms of QALYs may incentivize the adoption of seemingly eye wateringly expensive interventions. This was demonstrated in 2019, with the approval of Zolgensma for the treatment of Spinal Muscular Atrophy, at a cost of \$2.1million for a single dose – the most expensive drug ever successfully brought to market. This pricing was attributed to the massive healthcare benefit of successful therapy over an entire lifetime. It is not difficult to imagine the use of similar economic incentives to justify the often high cost of AI development. Once developed, such systems are relatively cheap to scale and deploy, enabling the potential for AI to diminish global health disparities.

Despite these challenges, pediatric AI is a nascent and growing field that presents substantial potential benefits to children and their families around the world. There are positive reasons to hope that, with appropriate incentives, support, and protections, pediatrics could become a pioneering specialty at the cutting edge of AI development.

Recent Developments in Pediatric AI

Cardiology

Despite an abundance of data of the type that is particularly well-suited to AI-based applications, the status on the use and/or development of AI-based tools in pediatric cardiology was described in one 2019 review as “dormant” [8]. However, there has been much more activity in adult cardiology, including automated or enhanced classification of cardiac images, electrocardiograms, and clinical prediction models [8, 9]. In many instances, the only limitation of their application to pediatric patients is the lack of research dedicated to cardiac conditions in children.

The diagnosis, management, and palliative and corrective surgery for children with congenital heart disease (CHD) is based on a thorough understanding of the complex anatomy related to a particular condition. Up to now, clinicians have largely had to understand this anatomy by visualizing two-dimensional (2D) images into a three-dimensional (3D) space. The use of virtual reality (VR) has begun to change this. Although this technology is in its infancy, the use of VR in which clinicians could view their patient’s entire heart (and even get inside of it, virtually) in an ultra-realistic 3D space could revolutionize the field, particularly in planning surgery on pediatric patients with cardiac disease [10]. Plasencia et al. have developed a methodology based on rendered 3D images to better match cardiac size for pediatric heart transplant patients with prospective donors [11]. Currently, the clinical standard for size matching is donor-recipient body weight (DRBW) ratio, which may unnecessarily limit a

patient’s access to a transplant. For the proposed methodology, computerized tomography (CT)/magnetic resonance imaging (MRI) is used to render a 3D image of the patient’s thorax. If there are available images from the donor, a 3D image of the donor’s heart can be rendered and be virtually placed in the donor’s chest. These investigators also developed a library of 90 hearts using CT/MRI data from pediatric patients of different sizes. This library could then be used to choose a representative heart for a donor in the likely scenario in which CT/MRI data on the donor’s heart are not available.

Segmentation of the different cardiac chambers is an important step toward providing structural information that facilitates the diagnosis of anatomical and functional cardiac disorders. Manual segmentation, which is the current gold standard, is time consuming and involves both significant inter- and intra-observer variability [12]. The task is hard in pediatric CHD due to significant variation in anatomy and a higher degree of motion artifacts in infants and children. To facilitate progress in CHD, a public competition for segmentation of CHD MRI data was introduced in 2016 [13]. The competition was won by Yu et al. with “3D FractalNet” for whole heart and great vessel segmentation [14]. A range of other successful approaches have been proposed since using deep learning architectures [15, 16] and more recently generative adversarial networks [17].

A number of studies have shown promise for machine learning algorithms assisting the diagnosis of heart murmurs using machine learning [18–20], often using an open access PhysioNet Challenge database of normal and abnormal heart sounds [21]. Tools that can distinguish between innocent and pathological murmurs could assist triage and reduce time to diagnosis, while also assisting pediatricians in remote locations without access to specialist pediatric cardiologists [22].

Despite the potential clinical benefits of enabling automated diagnosis of CHD from echocardiography images, there have been only limited studies conducted to date. The acquisition of high quality standard views is a key starting point for diagnosis, and there have been a number of studies demonstrating automated standard plane

detection for fetal echocardiography [23–25]. Wang et al. proposed a framework to classify atrial septal defects and ventricular septal defects from five-view echocardiogram videos, achieving an accuracy of 95.4% for CHD vs negative classes [26]. Le et al. developed a random forest-based framework to classify CHD from normal controls in fetal echocardiography, with accuracy of 0.91 [27]. Arnaout et al. described an ensemble of neural networks for distinguishing between normal hearts and complex CHD, achieving an area under the receiver operating characteristic curve (AUC) of 0.99, sensitivity of 95%, and specificity of 96% in a dataset from 4108 retrospective fetal echocardiograms, which is comparable to human performance [28].

In terms of risk prediction, Ruiz-Fernandez and colleagues developed four algorithms to predict postsurgical mortality using a dataset of over 2400 pediatric patients who underwent cardiac surgery for a variety of conditions, using multiple AI techniques [29]. The algorithms had an accuracy of 80–99% in predicting death. Clinical prediction models such as these could help inform clinicians in the future when deciding on the proper management of children with congenital heart disease.

Respiratory

Most investigations of AI-based applications for use in diagnosing and/or managing respiratory conditions have been focused on three areas: auscultation, interpretation of pulmonary function tests, and imaging [30]. As with cardiology, much of this work has been focused on adults. However, there are some unique aspects of pediatric respiratory medicine that provide opportunities for innovation that could substantially improve the health of children.

Worldwide, pneumonia remains a leading cause of morbidity and mortality in children under 5 years old. In many of the regions where the rate of childhood pneumonia is the highest, there is a scarcity of clinicians with the skill to accurately interpret pediatric chest radiographs. Thus, the use of AI to aid in the interpretation of

chest radiographs in children is of particular interest. AI-based interpretation of chest radiographs in young children is complicated by variations in thymic size and cardiothoracic ratio. Despite this, there have been significant advances in developing algorithms for diagnosing pneumonia using AI-based interpretation of chest radiographs in young patients. In a study that included chest radiographs on 858 South African children, an AI-based algorithm was used to classify each radiograph as pneumonia, “other infiltrate” or “no-infiltrate.” The ground truth for the reading was based on a consensus of three radiologists who independently reviewed, and used World Health Organization criteria to classify each radiograph; 333 (39%) of the images were classified as pneumonia by consensus. The AI-based algorithm had an AUC for correctly classifying a radiograph as pneumonia vs no pneumonia of 0.85, with a sensitivity of 76% and specificity of 80%. The results of this study highlight the potential of this technology, but also indicate its current limitations. An additional experienced radiologist interpreted a randomized sample of the study radiographs, stratified by ground truth classification, and had a sensitivity of 90% and specificity of 94% for identifying pneumonia, thus superior to the algorithm classification [31]. Another study by Naydenova et al. used support vector machines and random forests to predict pneumonia using temperature, respiratory rate, heart rate, and oxygen saturation [32]. In a test dataset of 1093 children with 777 pneumonia and 316 healthy controls, their approach achieved 96.6% sensitivity and 96.4% specificity. The simplicity of this approach means it could be embedded in smartphone decision support tools and used by basically trained healthcare workers in resource-constrained settings.

Lung ultrasound has proved to be a useful tool for detecting areas of consolidation as evidence of pneumonia, although it is operator dependent and requires interpretation by a skilled person. Correa et al. describe an approach using neural networks to classify lung ultrasound images into pneumonia vs normal lung groups, with 90.9% sensitivity and 100% sensitivity [33]. This work presents another opportunity to improve the challenge of diagnosis

of pneumonia in developing countries where resources are scarce.

Unlike adults, there are certain pediatric conditions that are characterized by highly specific cough characteristics, e.g., the “whoop” in pertussis, and the barking cough in croup. Parker et al. used hand-crafted features to classify coughs as pertussis or non-pertussis, achieving sensitivity of over 90% [34]. Sharan et al. used an iPhone to record coughs in children seen in clinical settings in Australia and developed an algorithm to identify patients with croup based on their cough acoustics [35]. After collecting data for a training sample, the algorithm was tested in 13 children with a clinical diagnosis of croup and 102 with a cough from other respiratory conditions. In this test sample the sensitivity of the algorithm in diagnosing croup was 92.3% and the specificity was 85.3%. Although there are weaknesses in this study, such as the lack of viral testing to help confirm a diagnosis of croup and the age of the patients (the mean age of the croup patients was >5 years in the test sample), these results highlight the potential for an AI-based approach that could be used by parents to determine if the likely etiology of their child’s cough was croup.

Wheezing is a primary sign of multiple pulmonary conditions, most notably asthma. Analysis of collected breath sounds in children has been shown to accurately document wheezing in noisy outpatient settings, in pediatric intensive care units and throughout the night [36–38]. However, special equipment is needed to collect the breath sound audio and the clinical applicability of the technology is uncertain. Because asthma is among the most common chronic diseases in pediatrics, there is a need for innovative technologies that would allow for passive and longitudinal monitoring of children in their home environment. Using an under the mattress ballistocardiographic sensor, Huffaker et al. assessed heart rate and heart rate variability, calculated respiratory rate, relative stroke volume, and assessed movement during the night in children with a history of asthma. These data were used to develop algorithms to determine when a child had asthma symptoms, predict the onset of these symptoms, and predict changes in asthma control test (ACT)

results (ACT is a validated questionnaire used to assess adequacy of asthma control). The algorithm had a sensitivity of 47% and a specificity of 96% in predicting patient-reported asthma symptoms and a sensitivity of 35% in predicting the onset of symptoms one day before they occurred. Changes in several of the physiologic parameters were statistically associated with changes in ACT scores [39].

Deep learning models have also been trained to predict a child’s asthma control up to a week ahead [40]. Using data collected from a weekly asthma self-monitoring tool, the best model achieved an accuracy of 71.8% and AUC of 0.757 when predicting the level of asthma control as “uncontrolled” (vs “controlled”), derived from a modified Asthma Control Test [41]. This study potentially supports future real-time decision support and personalized early warnings of asthma control deterioration.

Cough is a common symptom in children with asthma, and it has been suggested that nocturnal cough frequency may be a valid marker of asthma control. In a Japanese study, investigators collected audio data and movement of the abdomen with an accelerometer and developed an algorithm to objectively count coughs [39, 42]. The microphone and accelerometer were placed on children hospitalized with asthma and those with coughs from other respiratory conditions during an 8-h period at night. Patients with asthma had a mean of 144 ± 125 coughs per night compared to 18.3 ± 9 for those without asthma ($P < 0.001$). Those with an asthma exacerbation characterized as severe had significantly more coughs than those with a moderate exacerbation. Interestingly, patients with asthma had a distinct pattern of coughing with most coughs occurring shortly after falling asleep and shortly before awakening in the morning, a pattern not observed in children with coughs due to other etiologies. These results suggest that tracking nighttime cough counts, timing, and probably trajectory, could be used to identify a child with asthma and monitor severity over time. The obvious need is for an inexpensive AI-based system suitable for home use that can ideally evaluate coughing in a contactless manner.

Genetics

With the relative decline of both infectious disease and birth trauma in the Western world, congenital malformations have come to represent an increasing proportion of mortality and morbidity among children, particularly infants [43, 44]. A significant subset of such congenital malformations has a genetic basis, particularly in the NICU. Recent studies suggest underlying genetic causes for around 20% of NICU admissions [45], and a similar proportion of infant deaths [46] in tertiary centers. As such, rapid and accurate genetic diagnosis is particularly relevant in children and infants. There are three broad areas of genetic diagnosis at which AI systems show promise of significant clinical impact – phenotyping, interpretation of sequencing data, and prediction of disease risk from known genetic variants.

A first stage in genetic diagnosis, where no specific disease is known to run in the child's family history, is a thorough examination of the patient, and review of their clinical history – a process known as *phenotyping*. In some cases, this process may yield a confident genetic diagnosis prior to any confirmatory testing by a sufficiently experienced clinician. Image-based AI systems have been shown to perform at or above expert level in predicting syndromic genetic conditions from photographs of facial dysmorphology [47]. Other studies have used natural language processing to identify phenotypic features from the EHR and inform genetic testing [48].

Following phenotyping, an experienced clinician will decide on the exact form of genetic testing that is indicated, depending on the differential diagnosis. Sequencing methods, in particular, are increasing in popularity as the cost falls and speed increases. Currently, an entire genome can be sequenced in less than 1 day [49], at a cost of approximately \$700 in the USA in 2020 [50], although this will vary internationally. This low cost has resulted in a relative shift away from more targeted techniques such as gene panels, microarray, and karyotyping. For example, “virtual gene panels,” whereby the entire genome is sequenced, but only genes known to be linked to the patient’s phenotypic features are interpreted,

could become cheaper than traditional custom gene panels.

Unfortunately, the time and cost of interpreting the massive volume of information produced by whole-genome sequencing has not fallen at the same pace as the sequencing itself. Next generation sequencing in particular requires alignment and statistical interpretation in order to infer the presence of genetic variations. Furthermore, interpretation of such variants requires careful synthesis of clinical phenotyping, genetic information, and published case reports, since all human genomes will contain many variants that may or may not contribute to disease. It may take many days or weeks for specialist clinicians and clinical scientists to manually review each case and report results. In this context, work has focused on both variant calling [51] and variant pathogenicity interpretation [52], and deep learning approaches have outperformed traditional methods for identifying variants [53]. Further work has demonstrated the potential applicability of end-to-end AI systems that combine phenotypic inference from EHR data with variant interpretation, resulting in preliminary automated genetic reports from WGS within 24 h of sample collection [48].

Endocrinology

Type 1 diabetes, in which the pancreas produces insufficient or no insulin, has a typical age of onset between 4–7 years, with a second peak at 10–14 years. It is important to maintain glucose levels in a target range (typically 4–7 mmol/L). Significant time spent in a hyperglycemic state increases the risk of damage to a range of end-organs including the retina, kidneys, and peripheral nerves. However, hypoglycemia presents a more immediate danger of coma or even death. Therefore, tight and accurate control of blood sugar level is critical in ensuring safety and long-term health of patients. To this end, AI systems have been developed to achieve several interrelated goals.

Continuous glucose monitoring (CGM) systems [54, 55] measure glucose levels in the interstitial fluid, a close proxy for blood sugar level, at

a regular frequency on the order of 1–10 min intervals, providing richer monitoring than intermittent self-testing, which may only be conducted a few times a day. Since the advent and widespread adoption of CGM, there has been interest in using such data to monitor and adjust treatment regimens. This is typically done by uploading the data for clinical review at regular intervals, with physicians or specialist nurses advising on adjustments to insulin dosing. AI tools have been developed, which can provide such feedback, and show good agreement with endocrinologists in adult populations with weekly adjustment schedules [56], and have been demonstrated to be non-inferior to clinician advice in pediatric populations [57].

Taking the control system one step further, the development of an “artificial pancreas” [58] links the output of a CGM system with a continuous insulin delivery system, gated by onboard logic. This typically takes the form of a “fuzzy logic” system, which has been shown to be safe and effective in improving nocturnal euglycemia in pediatric patients [59], including during time spent away from the home setting [60]. Such systems have also been shown to be superior in terms of overall glycemic control, versus traditional approaches [61], reducing periods of hyperglycemia and hypoglycemia both overnight and with 24 h usage [62, 63]. Alternate control logic have been trialed, such as including reinforcement learning [64] and artificial neural networks, which have been shown to be effective *in silico* [65].

Beyond immediate control of blood sugar, AI systems have been developed to provide an “early warning” of predicted future deviations from normoglycemia up to 1–2 h prior to occurrence, using a wide variety of techniques [61].

Neonatology

Neonatology presents arguably one of the most promising domains for machine learning researchers in medicine. Preterm infants on the NICU generate a vast quantity of data [6], including types that are already known to be suitable for

AI techniques. EHRs are typically well coordinated, while large networks that aggregate data already exist in many countries (e.g., Vermont Oxford Network [66], the UK national neonatal research database [67]). Diseases are typically more homogeneous compared to other medical disciplines, with fewer confounders from complex past medical histories. Finally, earlier detection of deterioration and disease in neonatology can make a significant difference to outcome, and automated methods to flag actionable insights are likely to transform the practice of neonatology in the future.

Neonatal Sepsis

Among neonates born at >34 weeks gestation, early onset sepsis (EOS) is a rare event, occurring at a rate of 0.2–0.7 cases/1000 live births [68–70]. However, because of the potentially devastating consequences of failing to initiate antibiotic therapy early, many newborns are treated empirically while awaiting blood culture results. Based on criteria in national guidelines, it has been estimated that 5% of all newborns in the USA and 16% in the UK might receive empiric antibiotics because of various risk factors [68, 69].

Puopolo and colleagues analyzed a large dataset and used regression modeling and other mathematical techniques to develop a more efficient scoring system for assessing risk of EOS in a newborn [71]. The initial scoring system was based on objective perinatal information such as gestation, maternal temperature, group B streptococcal (GBS) screening status, and duration of rupture of membranes to provide a numeric estimate, i.e., the risk of sepsis, in a given baby. The group expanded the scoring system to include neonatal events including APGAR scores and vital signs [72]. Clinicians can access a website for the sepsis calculator (<https://neonatalsepsiscalculator.kaiserpermanente.org/>), enter the pertinent data, and instantly receive an estimate of the estimated risk of EOS in a particular infant. The authors of the calculator have recommended threshold scores for starting antibiotics [72].

Studies have been conducted that indicate a substantial reduction in use of unnecessary antibiotics when using the sepsis calculator compared

to use of national guidelines, without apparent increase in morbidity or mortality [68, 69]. Thus, there is evidence that a web-based tool, based on standard modeling techniques, can improve care. However, there have been reports of newborns with a low calculated sepsis risk score who developed EOS, highlighting the need for an AI-based scoring system that utilizes more objective data in the maternal and neonatal EHR [73]. The unique nature of the newborn health record, including frequent vital signs, limited provider notes, screening laboratory findings, and perinatal monitoring data, may be a particularly good example of how AI-based clinical prediction models can transform care. One envisions a sepsis risk score that is part of the newborn's EHR birth and is constantly updated with each new bit of information.

Jaundice

A central focus of newborn care is identifying infants with moderate hyperbilirubinemia so that treatment can be initiated, thus preventing rise to levels associated with development of acute or chronic bilirubin encephalopathy (i.e., kernicterus). During the newborn hospitalization in the USA and other high income countries, screening newborns for hyperbilirubinemia is typically done by obtaining blood for a total serum bilirubin (TSB) level, regardless of clinical condition, or measurement of transcutaneous bilirubin (TcB) level [74].

Jaundice screening is more complicated once a newborn is discharged [75]. Use of TcB meters is limited in the outpatient setting by the costs of these devices. Thus, most screening of jaundice in a newborn following discharge from the hospital is based on visual assessment, a technique that is prone to error [76, 77]. Clinical experience suggests there are at least three factors that limit the accuracy of visual assessment of neonatal jaundice. First, there is imperfect memory of what shade of yellow corresponds to a specific TSB level. Second, assessing jaundice in newborns with a variety of skin pigmentation is complicated. Finally, and perhaps most important, the appearance of jaundice varies based on the different lighting conditions commonly encountered in

clinical settings and it is virtually impossible to standardize lighting conditions.

All of these limitations are potentially solvable by an AI-based algorithm using data from digital images of a newborn's skin or sclerae. Several groups of investigators have reported on multiple systems, most based on using a commodity smartphone to capture images and software that includes machine-learning based algorithms to convert the digital information from the image(s) into an estimated bilirubin level. To account for lighting conditions these devices typically include a color calibration card that is included in the images, or add a small piece of hardware on the phone that is placed on the newborn's skin and use an external light source, usually the smartphone flash. Outlaw et al. reported on a system that attempted to eliminate issues related to skin tone and account for different lighting conditions [78]. First, images were captured of the infant's sclera instead of skin, thus eliminating the effects of melanin. Second, images both with "flash" and non-flash were obtained and the effect of ambient light "subtracted" out.

Most of the reported studies on the accuracy of smartphone-based devices to estimate neonatal jaundice are of modest sample size and frequently lack a diverse population of neonates. Correlation with the ground truth TSB levels are typically in the 0.5–0.91 range, although most investigators report high sensitivities and moderate specificities for identifying an infant with "hyperbilirubinemia," using a threshold TSB value to define hyperbilirubinemia [78–81]. Perhaps the most thoroughly investigated system has been developed by researchers at the University of Washington. In a racially diverse sample of 530 newborns (black infants comprised 20.8% of the study population and Asian infants 21.2%), they reported a correlation with TSB of 0.91; using two different recommended screening approaches the device had a sensitivity of 85% and 100%, respectively, for identifying a newborn with hyperbilirubinemia with corresponding specificities of 75% and 76% [79].

The main issue for this technology is how to be robust across different smartphone models (most studies have utilized a single model of phone) and

account for changes in photo processing that accompany changes in the operating system used by the phone. In addition, most research to date has been done in hospital or other clinical settings. It is largely unknown how well the developed algorithms will perform in other lighting environments, most notably in an infant's home.

In high income countries the use of a smartphone-based device to estimate bilirubin levels could assuage the fears of parents and healthcare providers when estimated levels are reassuringly low, and allow for more effective and earlier triage of patients who require phototherapy. More importantly, implementation of a low-cost device to identify newborns with significant jaundice could potentially save thousands of lives annually in low-income areas where bilirubin encephalopathy remains a major source of neonatal morbidity and mortality [82].

Ophthalmology

Ophthalmology is a specialty that is naturally suited to deep learning techniques, due to the widespread use of digital imaging modalities such as fundus photography, optical coherence tomography, and slit lamp imaging. This was demonstrated by early landmark success in adult AI-enabled diabetic retinopathy screening [83], which was able to directly build upon groundbreaking advances in computer vision performance made outside of the medical domain. In pediatric ophthalmology, a key application is in the detection and grading of retinopathy of prematurity (ROP). ROP mirrors many of the machine learning challenges of adult diabetic retinopathy, with significant variability in specialist accuracy of "plus" disease diagnosis, even among world-class experts (e.g., [84, 85]), with consequent difficulty in defining robust ground truth labels for model training. Plus disease in ROP describes the most severe vascular changes of dilation and tortuosity, and is associated with visual morbidity if left untreated. The potential benefits are significant – in developed countries, there are limited numbers of

ophthalmologists who feel comfortable managing ROP, either due to lack of training or medico-legal concerns; in developing countries, there is a lack of trained ophthalmologists, compounded by variable oxygen monitoring practices that lead to a higher incidence of severe ROP. As a result, the potential for AI tools to democratize access to subspecialist diagnostic skills, while reducing human variability, makes this a compelling area for future impact.

Early work in ROP was performed by Worrall et al. in 2016 [86], although the study was limited by a variable ground truth for plus disease. Brown et al. subsequently published a deep learning system called "i-ROP DL" trained on 5511 retinal photographs, classified into normal, pre-plus, and plus disease [87]. On an independent set of 100 images, for binary detection of plus disease, the system achieved sensitivity of 93% with specificity of 94%, with equivalent performance to human experts. More recently Hu et al. demonstrated a deep learning system that detects and grades ROP severity into five stages [88]. The system achieved an AUC of 0.992 for detection of ROP, and 0.921 for classification between mild/severe forms of ROP.

Longitudinal monitoring of ROP using quantitative measures of severity using deep learning may allow better prediction of risk and enable earlier treatment. Taylor et al. demonstrated that eyes that progress to treatment have a distinct progression of disease, and presented results showing possibility of predicting eyes at high risk 1 month before treatment [89].

Other applications have included detection of neonatal retinal hemorrhage and cataract detection. Wang et al. used 48,996 digital fundus images from 3770 newborns with retinal hemorrhages of different severities alongside normal controls, obtained from a large multicenter cross-sectional study performed in China [90]. A multi-class deep learning model was trained to classify scans into hemorrhage-free, grade 1, 2, or 3, with overall accuracy of 97.4%. Deep learning has also been used to classify pediatric cataracts from slit-lamp images, with sensitivity and specificity of 97.3% and 96.8%, respectively [91].

Primary Care

A range of interesting AI applications have been developed that may assist the primary care setting. For example, combining the ubiquitous smartphone with novel algorithms has enabled a new approach to middle ear infections. Chan et al. published a technique that uses a smartphone's speaker in combination with a paper funnel to send a "soft acoustic chirp into the ear canal," and employs a logistic regression machine learning algorithm to classify the reflections and predict middle ear fluid status [92]. In a leave-one-out cross-validation study using 98 patient ears at a pediatric surgical center, the authors obtained an AUC of 0.898, which compared favorably with commercial acoustic reflectometry (requiring custom hardware) with an AUC of 0.776. The system was described as being easily operated by children's parents without formal training, potentially aiding the future diagnosis of middle ear infections.

Another smartphone app has been developed that estimates hemoglobin levels by analyzing color and metadata of fingernail bed smartphone photos [93]. While not a dedicated pediatric study, subjects' ages ranged from 1 to 62 years old. Accuracy was shown to be $\pm 2.4 \text{ g dL}^{-1}$, with a sensitivity of 97% to detect anemia at a cutoff of 12.5 g dL^{-1} when compared with laboratory hemoglobin levels ($n = 100$ subjects). With additional calibration to a specific individual, accuracy was achieved at $\pm 0.92 \text{ g dL}^{-1}$. This technology offers the potential to improve the experience of care by offering needle-free alternatives to anemia monitoring for children with chronic conditions.

Novel uses of sensors may help make diagnoses in the future. In a study by Muñoz-Organero and colleagues, a convolutional neural network was designed to classify data from two tri-axial accelerometers (one on the wrist of the dominant arm, and the other on the ankle of the dominant leg) [94]. The study was small with 11 children with ADHD, and 11 controls, but demonstrated significant group differences, with classification accuracy of 87.5% for the wrist sensor (sensitivity = 0.6, specificity = 1), and 93.8% for the ankle sensor (sensitivity = 0.8, specificity = 1).

AI models to help triage and diagnose conditions may one day improve consistency and accuracy of pediatric care. A study from the Guangzhou Women and Children's Medical Center used an automated national language processing system using deep learning techniques to analyze a large EHR comprising 1,362,559 pediatric patient visits presenting to a major referral center [95]. The AI model was accurate when diagnosing conditions including upper respiratory infections, acute asthma exacerbations, sinusitis, and bacterial meningitis. When compared to two pediatricians split into five groups based upon proficiency and years of clinical experience, the model outperformed the two most junior physician groups, but underperformed the three most senior physician groups.

Radiology

Current AI techniques are naturally suited to imaging modalities, and a range of interesting applications have been developed in pediatrics, from classification and risk prediction, to technical aspects including improved image acquisition and reconstruction. Detection of imaging abnormalities has already been demonstrated in early work including pneumonia [96, 97], developmental dysplasia of the hip [97], wrist fractures [98], elbow effusions [99], and congenital urological abnormalities [100].

Estimation of skeletal maturity has been shown in multiple studies to be possible using deep learning approaches, with accuracy similar to or exceeding that of an expert radiologist [101–105]. Skeletal bone assessment is commonly performed in pediatrics for both diagnostic and therapeutic investigations, including endocrinology conditions, growth disorders, and genetic disorders. Bone age estimations are known to have significant intra- and interobserver variability [106, 107], making this an attractive and popular clinical problem to standardize through automation. Deep learning approaches appear to provide improvements vs previous feature-extraction techniques such as those utilized by commercial systems including one called Bone eXpert

[108]. In one well-designed study, Larson and colleagues trained a model using over 14,000 clinical hand radiographs, and tested using the publicly available Digital Hand Atlas dataset, with a root mean square error of 0.73 years.

In neonatal brain MRI, a deep learning model was used to classify dysplastic cerebelli in term infants born with congenital heart disease, achieving a classification accuracy of 0.985 for subtle cerebellar dysplasia [109]. Segmentation of neonatal brains is a historically difficult task given the rapid growth of the developing brain, with atlas-based techniques requiring multiple average atlases for each period of fetal and infant brain development. Atlas-based techniques are typically fragile, and may fail in the case of poor contrast, or with significant spatial variability compared to standard atlases – both of which are common in infant brain imaging. Deep learning techniques have enhanced accuracy with this task, including achieving robust hippocampal segmentation [110] and whole brain segmentation [111–114].

In pediatrics, fast imaging is essential to achieving high quality images with successful acquisition protocol completion before children become restless or claustrophobic. MRI techniques to improve speed of acquisition (i.e., [115, 116]) have the potential for significant utility in pediatric imaging, assuming accuracy can be assured. Minimization of radiation dose is also a key focus in pediatric patients, and techniques to reduce CT radiation dose show promise [117], alongside reductions in contrast agents required to achieve high quality MR scans [117, 118].

Pediatric Intensive Care

Pediatric intensive care (PICU) is a data-rich specialty, with many potential applications for AI. Areas explored so far include detection of sepsis [119], prediction of in-hospital mortality [120], prediction of need for transfer to PICU [121], and prediction of cardiac arrest [122].

Severe sepsis is a major clinical challenge in pediatrics, where earlier diagnosis and early goal-directed therapy have been shown to reduce mortality. Diagnosis is often delayed in hospitalized patients [123], and so automated accurate

methods to flag patients at high risk for sepsis would represent a major advance. A study by Kamaleswaran et al. used a range of methods including deep learning to predict the onset of severe sepsis using physiological markers in a dataset of 493 critically ill children [119]. They found that they could predict severe sepsis up to 8 h prior to a severe sepsis screening tool. Early work like this paves the way for future bedside monitors to be integrated with more intelligent tools to help clinicians anticipate the onset of sepsis through real time monitoring.

The potential for AI in PICU has arguably been neglected so far, with most AI research efforts so far focused on adult intensive care. Predicting outcomes and quantifying risk in children on the PICU is often more challenging than in the adult intensive care setting, largely due to the substantially changing physiology from prematurity to adulthood, but also smaller available datasets due to a reduced number of PICU beds than adult ICU beds. Next steps to advance the interoperability of EHR data, and effectively aggregate multicenter research databases will be required to create the scale necessary to train robust, clinically useful AI models in this setting.

Gastroenterology

There have been limited studies investigating AI methods in pediatric gastroenterology. Mossotto et al. described a system developed to classify pediatric inflammatory bowel disease, including Crohn's disease, ulcerative colitis, and inflammatory bowel disease unclassified (IBDU). The study used endoscopic and histological data from 287 children, and used a linear support vector machine to achieve classification accuracy of 83% of patients in an independent cohort of children [124].

Future Potential for AI in Pediatrics

AI presents the opportunity to transform aspects of pediatrics at a global scale, democratizing access to subspecialist diagnostic skills, improving quality and efficiency of care, enabling global

access to healthcare through sensor-rich internet-connected mobile devices, and enhancing imaging acquisition to reduce radiation while improving speed and quality.

Despite the potential for substantial health impact globally, a wide range of addressable clinical problems, and numerous compelling properties of pediatric datasets, AI research in pediatrics lags behind other adult specialties. Ongoing efforts to curate safe research-ready datasets to attract machine learning researchers and industrial developers will be essential to advancing the field.

Patient and carer acceptability of AI solutions is also yet to be properly addressed in pediatrics. While the public have been previously found to be overwhelmingly positive about the potential for AI to improve radiology [125], engagement with pediatricians, patients, and carers is essential to inform development of AI tools.

Over the past decade, it has become clear that AI systems can be trained to perform many tasks to a comparable or better level than human specialists. However, there are very few examples of AI systems actually deployed successfully in the real world [126]. The challenges facing the rest of medicine are unfortunately exacerbated in pediatrics, and the implementation gap between how these systems are developed and the setting where they will be deployed is a real challenge for the next decade. Robust evaluation through high quality clinical study design [127] and clear reporting [128] will be essential. The ultimate challenge will be for pediatricians to find ways to deploy these novel technologies into clinical practice in a manner that is safe, effective, equitable, and ultimately improves outcomes for children.

References

1. Sugiyama K, Hasegawa Y. Computer assisted medical diagnosis system for inborn errors of metabolism. *Jpn J Med Electron Biol Eng.* 1984;22:942–3.
2. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med.* 2019;25:30–6.
3. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25:44–56.
4. Alqahtani FF, Messina F, Offiah AC. Are semi-automated software program designed for adults accurate for the identification of vertebral fractures in children? *Eur Radiol.* 2019;29:6780–9.
5. Hoodbhoy Z, Masroor Jelani S, Aziz A, Habib MI, Iqbal B, Akmal W, Siddiqui K, Hasan B, Leeflang M, Das JK. Machine learning for child and adolescent health: a systematic review. *Pediatrics.* 2021. <https://doi.org/10.1542/peds.2020-011833>.
6. Khazaei H, Mench-Bressan N, McGregor C, Pugh JE. Health informatics for neonatal intensive care units: an analytical modeling perspective. *IEEE J Transl Eng Health Med.* 2015;3:3000109.
7. Nations U, United Nations. World population prospects 2019: highlights. Statistical Papers – United Nations (Ser A), Population and Vital Statistics Report. 2019. <https://doi.org/10.18356/13bf5476-en>.
8. Chang A. Artificial intelligence in pediatric cardiology and cardiac surgery: irrational hype or paradigm shift? *Ann Pediatr Cardiol.* 2019;12:191.
9. Gaffar S, Gearhart AS, Chang AC. The next frontier in pediatric cardiology: artificial intelligence. *Pediatr Clin N Am.* 2020;67:995–1009.
10. Sacks LD, Axelrod DM. Virtual reality in pediatric cardiology. *Curr Opin Cardiol.* 2020;35:37–41.
11. Plasencia JD, Kamarianakis Y, Ryan JR, et al. Alternative methods for virtual heart transplant-size matching for pediatric heart transplantation with and without donor medical images available. *Pediatr Transplant.* 2018;22:e13290.
12. Petitjean C, Dacher J-N. A review of segmentation methods in short axis cardiac MR images. *Med Image Anal.* 2011;15:169–84.
13. Pace DF, Dalca AV, Geva T, Powell AJ, Moghari MH, Golland P. Interactive whole-heart segmentation in congenital heart disease. *Med Image Comput Comput Assist Interv.* 2015;9351:80–8.
14. Yu L, Yang X, Qin J, Heng P-A. 3D FractalNet: dense volumetric segmentation for cardiovascular MRI volumes. In: Reconstruction, segmentation, and analysis of medical images. Springer International Publishing; 2017. p. 103–10.
15. Mukhopadhyay A. Total variation random forest: fully automatic MRI segmentation in congenital heart diseases. In: Reconstruction, segmentation, and analysis of medical images. Springer International Publishing; 2017. p. 165–71.
16. Pace DF, Dalca AV, Brosch T, Geva T, Powell AJ, Weese J, Moghari MH, Golland P. Iterative segmentation from limited training data: applications to congenital heart disease. *Deep Learn Med Image Anal Multimodal Learn Clin Decis Support.* 2018;11045:334–42.
17. Rezaei M, Yang H, Meinel C. Whole heart and great vessel segmentation with context-aware of generative adversarial networks. In: Bildverarbeitung für die Medizin 2018. Berlin/Heidelberg: Springer Vieweg; 2018. p. 353–8.
18. Bhatikar SR, DeGroff C, Mahajan RL. A classifier based on the artificial neural network approach for cardiologic auscultation in pediatrics. *Artif Intell Med.* 2005;33:251–60.

19. Latif S, Usman M, Rana R, Qadir J. Phonocardiographic sensing using deep learning for abnormal heartbeat detection. *IEEE Sensors J.* 2018;18:9393–400.
20. Kucharski D, Grochala D, Kajor M, Kańtoch E. A deep learning approach for valve defect recognition in heart acoustic signal. information systems architecture and technology. In: Proceedings of 38th international conference on information systems architecture and technology – ISAT 2017, 3–14. 2018.
21. Liu C, Springer D, Li Q, et al. An open access database for the evaluation of heart sound algorithms. *Physiol Meas.* 2016;37:2181–213.
22. Zühlke L, Myer L, Mayosi BM. The promise of computer-assisted auscultation in screening for structural heart disease and clinical teaching. *Cardiovasc J Afr.* 2012;23:405–8.
23. Baumgartner CF, Kamnitsas K, Matthew J, Fletcher TP, Smith S, Koch LM, Kainz B, Rueckert D. SonoNet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. *IEEE Trans Med Imaging.* 2017;36:2204–15.
24. Dong J, Liu S, Liao Y, Wen H, Lei B, Li S, Wang T. A generic quality control framework for fetal ultrasound cardiac four-chamber planes. *IEEE J Biomed Health Inform.* 2020;24:931–42.
25. Baumgartner CF, Kamnitsas K, Matthew J, Smith S, Kainz B, Rueckert D. Real-time standard scan plane detection and localisation in fetal ultrasound using fully convolutional neural networks. In: Medical image computing and computer-assisted intervention – MICCAI. Springer International Publishing; 2016. p. 203–11.
26. Wang J, Liu X, Wang F, Zheng L, Gao F, Zhang H, Zhang X, Xie W, Wang B. Automated interpretation of congenital heart disease from multi-view echocardiograms. *Med Image Anal.* 2021;69:101942.
27. Le TK, Truong V, Nguyen-Vo T-H, et al. Application of machine learning in screening of congenital heart diseases using fetal echocardiography. *J Am Coll Cardiol.* 2020;75:648.
28. Arnaout R, Curran L, Zhao Y, Levine JC, Chinn E, Moon-Grady AJ. Expert-level prenatal detection of complex congenital heart disease from screening ultrasound using deep learning. *medRxiv* 2020.06.22.20137786. 2020.
29. Ruiz-Fernández D, Monsalve Torra A, Soriano-Payá A, Marín-Alonso O, Triana Palencia E. Aid decision algorithms to estimate the risk in congenital heart surgery. *Comput Methods Prog Biomed.* 2016;126:118–27.
30. Ferrante G, Licari A, Marseglia GL, La Grutta S. Artificial intelligence as an emerging diagnostic approach in paediatric pulmonology. *Respirology.* 2020;25:1029–30.
31. Mahomed N, van Ginneken B, Philipsen RHJM, Melendez J, Moore DP, Moodley H, Sewchuran T, Mathew D, Madhi SA. Computer-aided diagnosis for World Health Organization-defined chest radiograph primary-endpoint pneumonia in children. *Pediatr Radiol.* 2020;50:482–91.
32. Naydenova E, Tsanás A, Casals-Pascual C, De Vos M. Smart diagnostic algorithms for automated detection of childhood pneumonia in resource-constrained settings. 2015 IEEE Global Humanitarian Technology Conference (GHTC). 2015. <https://doi.org/10.1109/ghtc.2015.7344000>.
33. Correa M, Zimic M, Barrientos F, et al. Automatic classification of pediatric pneumonia based on lung ultrasound pattern recognition. *PLoS One.* 2018;13: e0206410.
34. Parker D, Picone J, Harati A, Lu S, Jenkyns MH, Polgreen PM. Detecting paroxysmal coughing from pertussis cases using voice recognition technology. *PLoS One.* 2013;8:e82971.
35. Sharan RV, Abeyratne UR, Swarnkar VR, Porter P. Automatic croup diagnosis using cough sound recognition. *IEEE Trans Biomed Eng.* 2019;66:485–95.
36. Boner AL, Piacentini GL, Peroni DG, Irving CS, Goldstein D, Gavriely N, Godfrey S. Children with nocturnal asthma wheeze intermittently during sleep. *J Asthma.* 2010;47:290–4.
37. Habukawa C, Ohgami N, Matsumoto N, Hashino K, Asai K, Sato T, Murakami K. A wheeze recognition algorithm for practical implementation in children. *PLoS One.* 2020;15:e0240048.
38. Prodhan P, Dela Rosa RS, Shubina M, Haver KE, Matthews BD, Buck S, Kacmarek RM, Noviski NN. Wheeze detection in the pediatric intensive care unit: comparison among physician, nurses, respiratory therapists, and a computerized respiratory sound monitor. *Respir Care.* 2008;53:1304–9.
39. Huffaker MF, Carchia M, Harris BU, Kethman WC, Murphy TE, Sakarovich CCD, Qin F, Cornfield DN. Passive nocturnal physiologic monitoring enables early detection of exacerbations in children with asthma. A proof-of-concept study. *Am J Respir Crit Care Med.* 2018;198:320–8.
40. Luo G, Stone BL, Fassl B, Maloney CG, Gesteland PH, Yerram SR, Nkoy FL. Predicting asthma control deterioration in children. *BMC Med Inform Decis Mak.* 2015;15:84.
41. Nathan RA, Sorkness CA, Kosinski M, Schatz M, Li JT, Marcus P, Murray JJ, Pendergraft TB. Development of the asthma control test★A survey for assessing asthma control. *J Allergy Clin Immunol.* 2004;113:59–65.
42. Hirai K, Enseki M, Tabata H, Nukaga M, Matsuda S, Kato M, Furuya H, Mochizuki H. Objective measurement of frequency and pattern of nocturnal cough in children with asthma exacerbation. *Ann Allergy Asthma Immunol.* 2016;117:169–74.
43. Mathews TJ, Driscoll AK. Trends in Infant Mortality in the United States, 2005–2014. *NCHS Data Brief.* 2017;279:1–8.
44. Kyu HH, Stein CE, Boschi Pinto C, et al. Causes of death among children aged 5–14 years in the WHO European Region: a systematic analysis for the Global

- Burden of Disease Study 2016. *Lancet Child Adolesc Health.* 2018;2:321–37.
45. French CE, Delon I, Dolling H, et al. Whole genome sequencing reveals that genetic conditions are frequent in intensively ill children. *Intensive Care Med.* 2019;45:627–36.
46. Wojcik MH, Schwartz TS, Thiele KE, et al. Infant mortality: the contribution of genetic disorders. *J Perinatol.* 2019;39:1611–9.
47. Gurovich Y, Hanani Y, Bar O, et al. Identifying facial phenotypes of genetic disorders using deep learning. *Nat Med.* 2019;25:60–4.
48. Clark MM, Hildreth A, Batalov S, et al. Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. *Sci Transl Med.* 2019. <https://doi.org/10.1126/scitranslmed.aat6177>.
49. Genome Sequencing. 2015. <https://www.genomicsengland.co.uk/understanding-genomics/genome-sequencing/>. Accessed 15 Mar 2021.
50. National Human Genome Research Institute DNA Sequencing Costs: Data. <https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data>. Accessed 15 Mar 2021.
51. Poplin R, Chang P-C, Alexander D, et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol.* 2018;36:983–7.
52. Sundaram L, Gao H, Padigepati SR, et al. Predicting the clinical impact of human mutation with deep neural networks. *Nat Genet.* 2018;50:1161–70.
53. Supernat A, Vidarsson OV, Steen VM, Stokowy T. Comparison of three variant callers for human whole genome sequencing. *Sci Rep.* 2018;8:17851.
54. Mastrototaro JJ. The MiniMed continuous glucose monitoring system. *Diabetes Technol Ther.* 2000;2 (Suppl 1):S13–8.
55. Bode BW. Clinical utility of the continuous glucose monitoring system. *Diabetes Technol Ther.* 2000;2 (Suppl 1):S35–41.
56. Tyler NS, Mosquera-Lopez CM, Wilson LM, et al. An artificial intelligence decision support system for the management of type 1 diabetes. *Nat Metabolism.* 2020;2:612–9.
57. Nimri R, NextDREAM Consortium, Battelino T, Laffel LM, Slover RH, Schatz D, Weinzimer SA, Dovc K, Danne T, Phillip M. Insulin dose optimization using an automated artificial intelligence-based decision support system in youths with type 1 diabetes. *Nat Med.* 2020;26:1380–4.
58. Elleri D, Dunger DB, Hovorka R. Closed-loop insulin delivery for treatment of type 1 diabetes. *BMC Med.* 2011;9:120.
59. Elleri D, Allen JM, Biagioli M, et al. Evaluation of a portable ambulatory prototype for automated overnight closed-loop insulin delivery in young people with type 1 diabetes. *Pediatr Diabetes.* 2012;13:449–53.
60. Phillip M, Battelino T, Atlas E, et al. Nocturnal glucose control with an artificial pancreas at a diabetes camp. *N Engl J Med.* 2013;368:824–33.
61. Contreras I, Vehi J. Artificial intelligence for diabetes management and decision support: literature review. *J Med Internet Res.* 2018;20:e10775.
62. Bekiari E, Kitsios K, Thabit H, Tauschmann M, Athanasiadou E, Karagiannis T, Haidich A-B, Hovorka R, Tsapas A. Artificial pancreas treatment for outpatients with type 1 diabetes: systematic review and meta-analysis. *BMJ.* 2018;361:k1310.
63. Breton MD, Beck RW, Wadwa RP, iDCL Trial Research Group. A randomized trial of closed-loop control in children with type 1 diabetes. *Reply. N Engl J Med.* 2020;383:2484.
64. Bothe MK, Dickens L, Reichel K, Tellmann A, Ellger B, Westphal M, Faisal AA. The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas. *Expert Rev Med Devices.* 2013;10:661–73.
65. de Canete JF, Gonzalez-Perez S, Ramos-Diaz JC. Artificial neural networks for closed loop control of in silico and ad hoc type 1 diabetes. *Comput Methods Prog Biomed.* 2012;106:55–66.
66. Edwards EM, Ehret DEY, Soll RF, Horbar JD. Vermont Oxford Network: a worldwide learning community. *Transl Pediatr.* 2019;8:182–92.
67. Modi N. Information technology infrastructure, quality improvement and research: the UK National Neonatal Research Database. *Transl Pediatr.* 2019;8:193–8.
68. Goel N, Shrestha S, Smith R, et al. Screening for early onset neonatal sepsis: NICE guidance-based practice versus projected application of the Kaiser Permanente sepsis risk calculator in the UK population. *Arch Dis Child Fetal Neonatal Ed.* 2020;105:118–22.
69. Kuzniewicz MW, Puopolo KM, Fischer A, Walsh EM, Li S, Newman TB, Kipnis P, Escobar GJ. A quantitative, risk-based approach to the management of neonatal early-onset sepsis. *JAMA Pediatr.* 2017;171:365–71.
70. Cailes B, Kortsalioudaki C, Buttery J, Pattnayak S, Greenough A, Matthes J, Bedford Russell A, Kennea N, Heath PT, neonIN network. Epidemiology of UK neonatal infections: the neonIN infection surveillance network. *Arch Dis Child Fetal Neonatal Ed.* 2018;103:F547–53.
71. Puopolo KM, Draper D, Wi S, Newman TB, Zupancic J, Lieberman E, Smith M, Escobar GJ. Estimating the probability of neonatal early-onset infection on the basis of maternal risk factors. *Pediatrics.* 2011;128:e1155–63.
72. Escobar GJ, Puopolo KM, Wi S, Turk BJ, Kuzniewicz MW, Walsh EM, Newman TB, Zupancic J, Lieberman E, Draper D. Stratification of risk of early-onset sepsis in newborns \geq 34 weeks' gestation. *Pediatrics.* 2014;133:30–6.
73. Pettinger KJ, Mayers K, McKechnie L, Phillips B. Sensitivity of the Kaiser Permanente early-onset sepsis calculator: a systematic review and meta-analysis. *EClin Med.* 2020;19:100227.
74. Taylor JA, Burgos AE, Flaherman V, Chung EK, Simpson EA, Goyal NK, Von Kohorn I,

- Dhepyaswan N, Better Outcomes through Research for Newborns Network. Discrepancies between transcutaneous and serum bilirubin measurements. *Pediatrics*. 2015;135:224–31.
75. Maisels MJ, Bhutani VK, Bogen D, Newman TB, Stark AR, Watchko JF. Hyperbilirubinemia in the newborn infant >=35 weeks' gestation: an update with clarifications. *Pediatrics*. 2009;124:1193–8.
 76. Moyer VA, Ahn C, Sneed S. Accuracy of clinical judgment in neonatal jaundice. *Arch Pediatr Adolesc Med*. 2000;154:391–4.
 77. NICE. Neonatal jaundice – clinical guideline. CG98. 2010. <https://www.nice.org.uk/guidance/cg98/evidence/full-guideline-pdf-245411821>
 78. Outlaw F, Nixon M, Odeyemi O, MacDonald LW, Meek J, Leung TS. Smartphone screening for neonatal jaundice via ambient-subtracted sclera chromaticity. *PLoS One*. 2020;15:e0216970.
 79. Taylor JA, Stout JW, de Greef L, et al. Use of a smartphone app to assess neonatal jaundice. *Pediatrics*. 2017. <https://doi.org/10.1542/peds.2017-0312>.
 80. Rizvi MR, Alaskar FM, Albaradie RS, Rizvi NF, Al-Abdulwahab K. A novel non-invasive technique of measuring bilirubin levels using bilicapture. *Oman Med J*. 2019;34:26–33.
 81. Munkholm SB, Krøgholt T, Ebbesen F, Szecsi PB, Kristensen SR. The smartphone camera as a potential method for transcutaneous bilirubin measurement. *PLoS One*. 2018;13:e0197938.
 82. Bhutani VK, Zipursky A, Blencowe H, et al. Neonatal hyperbilirubinemia and Rhesus disease of the newborn: incidence and impairment estimates for 2010 at regional and global levels. *Pediatr Res*. 2013;74 (Suppl 1):86–100.
 83. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316:2402–10.
 84. Chiang MF, Jiang L, Gelman R, Du YE, Flynn JT. Interexpert agreement of plus disease diagnosis in retinopathy of prematurity. *Arch Ophthalmol*. 2007;125:875–80.
 85. Fleck BW, BOOST II Retinal Image Digital Analysis (RIDA) Group, Williams C, et al. An international comparison of retinopathy of prematurity grading performance within the Benefits of Oxygen Saturation Targeting II trials. *Eye*. 2018;32:74–80.
 86. Worrall DE, Wilson CM, Brostow GJ. Automated retinopathy of prematurity case detection with convolutional neural networks. In: Deep learning and data labeling for medical applications. Springer International Publishing; 2016. p. 68–76.
 87. Brown JM, Campbell JP, Beers A, et al. Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. *JAMA Ophthalmol*. 2018;136:803–10.
 88. Hu J, Chen Y, Zhong J, Ju R, Yi Z. Automated analysis for retinopathy of prematurity by deep neural networks. *IEEE Trans Med Imaging*. 2019;38:269–79.
 89. Taylor S, Brown JM, Gupta K, et al. Monitoring disease progression with a quantitative severity scale for retinopathy of prematurity using deep learning. *JAMA Ophthalmol*. 2019. <https://doi.org/10.1001/jamaophthalmol.2019.2433>.
 90. Wang B, Xiao L, Liu Y, Wang J, Liu B, Li T, Ma X, Zhao Y. Application of a deep convolutional neural network in the diagnosis of neonatal ocular fundus hemorrhage. *Biosci Rep*. 2018. <https://doi.org/10.1042/bsr20180497>.
 91. Liu X, Jiang J, Zhang K, et al. Localization and diagnosis framework for pediatric cataracts based on slit-lamp images using deep features of a convolutional neural network. *PLoS One*. 2017;12:e0168606.
 92. Chan J, Raju S, Nandakumar R, Bly R, Gollakota S. Detecting middle ear fluid using smartphones. *Sci Transl Med*. 2019;11:eaav1102. <https://doi.org/10.1126/scitranslmed.aav1102>.
 93. Mannino RG, Myers DR, Tyburski EA, Caruso C, Boudreaux J, Leong T, Clifford GD, Lam WA. Smartphone app for non-invasive detection of anemia using only patient-sourced photos. *Nat Commun*. 2018;9:4924.
 94. Muñoz-Organero M, Powell L, Heller B, Harpin V, Parker J. Using recurrent neural networks to compare movement patterns in ADHD and normally developing children based on acceleration signals from the wrist and ankle. *Sensors*. 2019;19:2935. <https://doi.org/10.3390/s19132935>.
 95. Liang H, Tsui BY, Ni H, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat Med*. 2019;25:433–8.
 96. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med*. 2018;15:e1002686.
 97. Li Q, Zhong L, Huang H, et al. Auxiliary diagnosis of developmental dysplasia of the hip by automated detection of Sharp's angle on standardized anteroposterior pelvic radiographs. *Medicine*. 2019;98: e18500.
 98. Thian YL, Li Y, Jagmohan P, Sia D, Chan VEY, Tan RT. Convolutional neural networks for automated fracture detection and localization on wrist radiographs. *Radiol: Artif Intell*. 2019;1:e180001.
 99. England JR, Gross JS, White EA, Patel DB, England JT, Cheng PM. Detection of traumatic pediatric elbow joint effusion using a deep convolutional neural network. *Am J Roentgenol*. 2018;211:1361–8.
 100. Zheng Q, Furth SL, Tasian GE, Fan Y. Computer-aided diagnosis of congenital abnormalities of the kidney and urinary tract in children based on ultrasound imaging data by integrating texture image features and deep transfer learning image features. *J Pediatr Urol*. 2019;15:75.e1–7.
 101. Tong C, Liang B, Li J, Zheng Z. A deep automated skeletal bone age assessment model with heterogeneous features learning. *J Med Syst*. 2018;42:249.
 102. Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. *Radiology*. 2018;287:313–22.

103. Mutasa S, Chang PD, Ruzal-Shapiro C, Ayyala R. MABAL: a novel deep-learning architecture for machine-assisted bone age labeling. *J Digit Imaging*. 2018;31:513–9.
104. Spampinato C, Palazzo S, Giordano D, Aldinucci M, Leonardi R. Deep learning for automated skeletal bone age assessment in X-ray images. *Med Image Anal*. 2017;36:41–51.
105. Kim JR, Shim WH, Yoon HM, Hong SH, Lee JS, Cho YA, Kim S. Computerized bone age estimation using deep learning based program: evaluation of the accuracy and efficiency. *AJR Am J Roentgenol*. 2017;209: 1374–80.
106. Thodberg HH, Sävendahl L. Validation and reference values of automated bone age determination for four ethnicities. *Acad Radiol*. 2010;17:1425–32.
107. Berst MJ, Dolan L, Bogdanowicz MM, Stevens MA, Chow S, Brandser EA. Effect of knowledge of chronologic age on the variability of pediatric bone age determined using the Greulich and Pyle standards. *Am J Roentgenol*. 2001;176:507–10.
108. Thodberg HH, Kreiborg S, Juul A, Pedersen KD. The BoneXpert method for automated determination of skeletal maturity. *IEEE Trans Med Imaging*. 2009;28:52–66.
109. Ceschin R, Zahner A, Reynolds W, Gaesser J, Zuccoli G, Lo CW, Gopalakrishnan V, Panigrahy A. A computational framework for the detection of subcortical brain dysmaturity in neonatal MRI using 3D convolutional neural networks. *NeuroImage*. 2018;178:183–97.
110. Guo Y, Wu G, Commander LA, Szary S, Jewells V, Lin W, Shent D. Segmenting hippocampus from infant brains by sparse patch matching with deep-learned features. *Med Image Comput Comput Assist Interv*. 2014;17:308–15.
111. Dolz J, Desrosiers C, Wang L, Yuan J, Shen D, Ben Ayed I. Deep CNN ensembles and suggestive annotations for infant brain MRI segmentation. *Comput Med Imaging Graph*. 2020;79:101660.
112. Zhang W, Li R, Deng H, Wang L, Lin W, Ji S, Shen D. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *NeuroImage*. 2015;108:214–24.
113. Moeskops P, Viergever MA, Mendrik AM, de Vries LS, Benders MJNL, Igum I. Automatic segmentation of MR brain images with a convolutional neural network. *IEEE Trans Med Imaging*. 2016;35:1252–61.
114. Nie D, Wang L, Gao Y, Shen D. Fully convolutional networks for multi-modality isointense infant brain image segmentation. *Proc IEEE Int Symp Biomed Imaging*. 2016;2016:1342–5.
115. Nguyen XV, Oztek MA, Nelakurti DD, Brunnquell CL, Mossa-Basha M, Haynor DR, Prevedello LM. Applying artificial intelligence to mitigate effects of patient motion or other complicating factors on image quality. *Top Magn Reson Imaging*. 2020;29:175–80.
116. Wang S, Su Z, Ying L, Peng X, Zhu S, Liang F, Feng D, Liang D. Accelerating magnetic resonance imaging via deep learning. *Proc IEEE Int Symp Biomed Imaging*. 2016;2016:514–7.
117. Chen H, Zhang Y, Kalra MK, Lin F, Chen Y, Liao P, Zhou J, Wang G. Low-dose CT with a residual encoder-decoder convolutional neural network. *IEEE Trans Med Imaging*. 2017;36:2524–35.
118. Gong E, Pauly JM, Wintermark M, Zaharchuk G. Deep learning enables reduced gadolinium dose for contrast-enhanced brain MRI. *J Magn Reson Imaging*. 2018;48:330–40.
119. Kamaleswaran R, Akbilic O, Hallman MA, West AN, Davis RL, Shah SH. Applying artificial intelligence to identify biomarkers predicting severe sepsis in the PICU. *Pediatr Crit Care Med*. 2018;19: e495–503.
120. Fernández IS, Sansevere AJ, Gaínza-Lein M, Kapur K, Loddenkemper T. Machine learning for outcome prediction in electroencephalograph (EEG)-monitored children in the intensive care unit. *J Child Neurol*. 2018;33:546–53.
121. Zhai H, Brady P, Li Q, Lingren T, Ni Y, Wheeler DS, Solti I. Developing and evaluating a machine learning based algorithm to predict the need of pediatric intensive care unit transfer for newly hospitalized children. *Resuscitation*. 2014;85:1065–71.
122. Kennedy CE, Aoki N, Mariscalco M, Turley JP. Using time series analysis to predict cardiac arrest in a PICU. *Pediatr Crit Care Med*. 2015;16:e332–9.
123. Weiss SL, for the SPROUT Study Investigators and Pediatric Acute Lung Injury and Sepsis Investigators (PALISI) Network, Fitzgerald JC, et al. Discordant identification of pediatric severe sepsis by research and clinical definitions in the SPROUT international point prevalence study. *Crit Care*. 2015;19:325. <https://doi.org/10.1186/s13054-015-1055-x>.
124. Mossotto E, Ashton JJ, Coelho T, Beattie RM, MacArthur BD, Ennis S. Classification of paediatric inflammatory bowel disease using machine learning. *Sci Rep*. 2017;7:2427.
125. Goldberg JE, Rosenkrantz AB. Artificial intelligence and radiology: a social media perspective. *Curr Probl Diagn Radiol*. 2019;48:308–11.
126. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. 2019;17:195.
127. Rivera SC, The SPIRIT-AI and CONSORT-AI Working Group, Liu X, Chan A-W, Denniston AK, Calvert MJ, SPIRIT-AI and CONSORT-AI Steering Group, SPIRIT-AI and CONSORT-AI Consensus Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med*. 2020;26:1351–63.
128. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26:1364–74.



AIM in Neonatal and Pediatric Intensive Care

75

David Forsberg, Antoine Honoré, Kerstin Jost, Emma Persad,
Karen Coste, Saikat Chatterjee, Susanne Rautiainen, and
Eric Herlenius

Contents

Introduction	1048
Sepsis Definition	1049
Continuous Vital Sign Assessment to Predict Life-Threatening Events	1049
Neonatal Sepsis and the NICU	1050
Pediatric Sepsis and Early Detection	1052
Challenges and Future Perspectives of Automated Vital Signs Pattern Analysis	1052
Conclusions	1053
References	1053

D. Forsberg · K. Jost · S. Rautiainen · E. Herlenius (✉)
Department of Women's & Children's Health, Karolinska Institutet, Stockholm, Sweden

Astrid Lindgren Children's Hospital, Children's and Women's Health, Karolinska University Hospital, Stockholm, Sweden
e-mail: Eric.Herlenius@ki.se

A. Honoré
Department of Women's & Children's Health, Karolinska Institutet, Stockholm, Sweden

Astrid Lindgren Children's Hospital, Children's and Women's Health, Karolinska University Hospital, Stockholm, Sweden

Division of Information Science and Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

E. Persad
Department of Women's & Children's Health, Karolinska Institutet, Stockholm, Sweden

Astrid Lindgren Children's Hospital, Children's and Women's Health, Karolinska University Hospital, Stockholm, Sweden

Karl Landsteiner University of Health Sciences, Krems, Austria

K. Coste
Department of Women's & Children's Health, Karolinska Institutet, Stockholm, Sweden

Astrid Lindgren Children's Hospital, Children's and Women's Health, Karolinska University Hospital, Stockholm, Sweden

CNRS, INSERM, GReD, Université Clermont Auvergne, Clermont-Ferrand, France

S. Chatterjee
Division of Information Science and Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

Abstract

Infections are one of the leading causes of death in infants and detecting life-threatening events in infants is challenging. Thus, providing effective life-saving interventions in time is essential. As infants' immune and autonomic control system are under development, signs preceding potentially life-threatening events are subtle. Clinical detection is aided by analysis of biomarkers, which unfortunately requires invasive sampling and is time consuming. Infection and inflammation interfere with the autonomic control systems and consequently affect vital signs. Constantly monitoring vital signs at a high frequency enables the immediate detection of discrepancies and is thus a key, noninvasive instrument in modern intensive care units. For pediatric intensive care, several predictive monitoring systems have been developed over the last decade *that aim to* utilize vital sign monitoring to mitigate the risk of developing life-threatening events, such as sepsis. Recent advances in the field of machine learning have provided novel techniques for big data analysis. This enables an individualized risk assessment via continuous multimodal inputs and development of better clinical decision support systems. These more advanced systems are able to detect sepsis 24 hours earlier than clinical practice and enable an overall risk assessment for future sepsis, life-threatening events, and death at the time of hospitalization or during the first week of life.

This chapter summarizes the current evidence on machine learning-based monitoring systems and provides an overview on the strengths, limitations, and potential future roles of novel machine learning-based methods for the early detection of pediatric sepsis and potentially life-threatening events.

Introduction

The prediction of life-threatening events among infants remains a major challenge for health care. Infants are particularly fragile and at risk of rapidly

developing cardiorespiratory dysfunction and sepsis, and during the first week of life, they are at higher risk of life-threatening events, such as apnea, bradycardia, and oxygen desaturation [1–4]. Infants affected by such events are typically admitted to a neonatal (new-born and premature infants) or pediatric (infants typically older than 3 days) intensive care unit (NICU/PICU). This unfortunately puts infants at an increased risk of infection, leading to increased morbidity and mortality and contributing to long-term complications, such as chronic lung disease and adverse neurodevelopmental outcomes [5–7]. Numbers based on UK pediatric population report an incidence of neonatal infections of 6.1/1000 live births, compared with 48.8/1000 among NICU patients [8].

The infections, including sepsis, pneumonia, and gastrointestinal infections, are leading causes of death [9, 10] as they are hard to diagnose, especially early in life due in part to the immaturity of the innate immune system. Early, and initially subtle, signs of infections include fluctuating temperature, reduced activity, reduced appetite, and a change in behavioral patterns [11, 12]. However, in the last decade it has become evident that, sepsis- and inflammation-related changes can also be identified in vital signs, such as heart and respiratory rate patterns [13].

A clinical assessment of vital signs could become essential for the early diagnosis of life-threatening events. However, as it requires a constant monitoring of the infant, which cannot be interpreted manually, it poses a new challenge in medicine [14]. Thus, during the last decades, a realization has emerged that it is crucial to develop effective predictive monitoring (Clinical Decision Support – CDS) systems for early detection of infection and possible life-threatening events. This will enable earlier and more efficient treatment, thereby reducing morbidity and mortality [13].

The present chapter aims to review the field of predictive monitoring systems for NICU and PICU settings and evaluate pros and cons of different approaches. It is an emerging and currently expanding field. Therefore, it is important to identify both the technical and clinical aspects that must be considered when employing a predictive monitoring system in everyday healthcare.

Sepsis Definition

Based on the Sequential [Sepsis-related] Organ Failure Assessment (SOFA) score, the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) has established a clear definition of sepsis and septic shock for adult patients [15]. These definitions have been used for pediatric sepsis as well, but are not optimal as age-related differences in pathophysiology and clinical manifestations needs to be taken into account, especially when assessing SOFA score [16], and for neonatal sepsis they are not applicable [17]. In 2019, the Society of Critical Care Medicine (SCCM) convened the Pediatric Sepsis Definition Task Force, to develop pediatric sepsis definitions and identify risk factors. This work is currently in progress and a first systemic review is under preparation [18]. Neonatal sepsis diagnosis is most often based on the isolation of a microbe in blood or cerebrospinal fluid [12], sometimes in combination with various biomarkers [19–21]. However blood cultures are often negative for microbes despite convincing clinical signs among neonates [22] and biomarker reference intervals vary among studies [17]. Thus, culture-negative sepsis is one of the most common diagnoses in preterm infants, leading to empiric use of antibiotics [23]. Clinical signs of neonatal sepsis are often due to a systemic inflammatory response and include tachycardia, tachypnea, fever, or hypothermia [12, 24].

Infections during the first months of life may dramatically aggravate underlying cardiorespiratory dysfunction [25]. Thus, vital signs monitoring is a crucial feature when developing novel CDS systems for sepsis risk stratification. Circulatory and respiratory responses are closely connected under the control of autonomic brainstem networks and the sympathetic and parasympathetic nervous system activity [26] and regulate heart rate, vascular resistance, and blood pressure [25, 27, 28]. Infections generate an immune system response leading to the release of various cytokines and signaling molecules [29], thereby inducing the production of prostaglandin E₂ (PGE₂) [30] and ultimately affecting autonomic control systems [30–32] and neuronal

behavior [33–35]. This leads to disruptions in cardiorespiratory rhythm in neonates [3, 32, 35–38] and affects the response of the cardiorespiratory system to hypoxia, anoxia, and hypercapnia [39], apparent in the increased prevalence of apnea, bradycardia, and desaturation events in neonates and young children [3, 25, 32].

Continuous Vital Sign Assessment to Predict Life-Threatening Events

Currently, many clinical decisions are based on the observations of vital signs annotated by clinical staff, who at best may report one average value per hour [40, 41]. In contrast, monitoring systems have the potential to record vital sign data every 8 milliseconds, providing over 75 million data points per week per monitored variable. The utilization of machine learning (ML) methods is thus advantageous in clinical practice as it allows for the recognition of discrete subtle changes in the vital signs, a feat that is impossible for humans.

It is now well-established that ABD events often precede life-threatening events in the NICU and PICU [2–4, 42–45]. Therefore, high-quality detection of small subtle alterations in vital sign patterns could improve the diagnostics of infections in NICU/PICU patients (Fig. 1) [46]. Modern NICUs and PICUs are equipped with advanced monitoring systems that continuously measure vital signs, such as heart rate, respiratory rate, oxygen saturation, and sometimes also blood pressure and body temperature. However, most current alarm systems are based on threshold values, e.g., saturation below 92% elicits an alarm. Utilizing ML algorithms could allow the detection of deviations in the continuous time series and generate risk scores for patient deterioration that can be used as CDS systems.

However, while many CDS systems are currently evaluated academically [13, 43–45, 47] few have been pushed to the open market. Additionally, the available systems have limitations imposed by restricted study populations and inclusion criteria in clinical trials, and thus the clinical implementation have been limited.

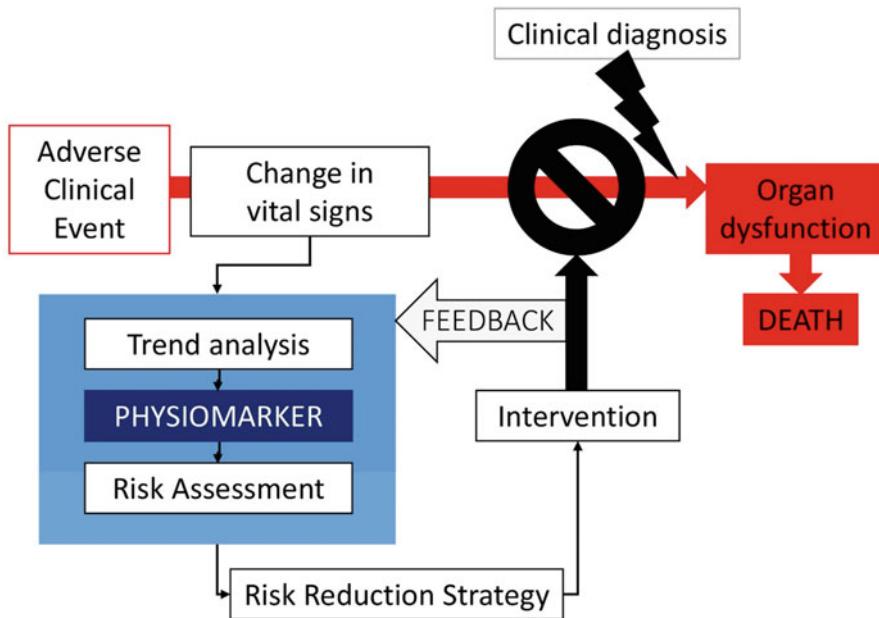


Fig. 1 Adverse events, that if untreated may lead to substantial morbidity or death, induce early, initially subtle changes in vital signs. These can be detected through a human in the loop machine learning-based trend analysis and from alterations in biomarkers for disease,

including sepsis and inflammatory-induced adverse events. A biomarker based alert can then be used for early risk reduction strategies or interventions to reduce morbidity and mortality

Moreover, data monitoring issues and high rates of false alarms may induce alarm fatigue and many clinicians hesitate to alter established routines without trust and interpretability of CDS systems. Nevertheless, hospital monitoring systems generate a huge amount of data and may detect subtle changes and trends, which would otherwise be overlooked by humans. Among the available CDS systems, the HeRO system is the most extensively evaluated and also commercially available for continuous patient evaluation and late-onset sepsis (LOS) risk assessment for neonatal patients [47].

Neonatal Sepsis and the NICU

In the NICU, heart rate characteristics (HRC) monitoring to detect sepsis has been in use for a decade. The commercial HeRO system was the first early warning system for sepsis in the NICU utilizing the continuous monitoring of electrocardiogram signals and HRC [48, 49]. It uses a time

window of input signals that is linked to an outcome of interest. By adapting a frame-by-frame approach, a score is computed using all measurements within a certain time window of a single variable. For example, the HeRO system utilizes 4096 heart beats per time window, corresponding to 20–30 min of heart beat RR intervals, and renders a score every hour, which is the risk of infection in the next 24 h [50]. The HeRO system has been implemented to predict *late-onset sepsis within 24 h* or *no late-onset sepsis within 24 h*. Used in very low birth weight infants, HeRO monitoring is effective in detecting sepsis [42, 49, 51] and lowers mortality [50]. A combination of HRC index and pulse oximetry during the first week of life can identify patients with a higher risk of developing sepsis or death sometime during their NICU stay [51]. Notably, patient demographics alone, including birth weight, gestational age, sex, antenatal steroids, and Apgar score provided an equivalent risk assessment [51].

In a comparison of 22 separate vital sign features for neonatal sepsis detection, the usefulness

of heart rate variability-based features, such as HRC, was confirmed [45]. Notably, features based on respiration and respiratory instability consistently outperformed features based on respiratory cessations, including cessations constituting clinical apneas and sometimes outperformed various features of HRC [45]. However, the authors emphasized the issue with impedance-based measurements of respiration using ECG electrodes. Additionally, each feature was compared separately while combining features did not improve the classifier-performance, as measured by AUROC, in the 49 patients included in their analysis, even though no data is shown. Combining features, e.g., HRC, respiration, oxygen saturation, and patient demographics have previously been shown to improve performance effectively [43, 44, 51].

A manually designed decision tree model has been developed to predict late-onset sepsis and necrotizing enterocolitis in $n < 32$ week preterm babies – the RALIS algorithm [43]. This algorithm utilizes multiple vital signs; heart rate,

respiratory rate, and temperature from monitoring devices and manually recorded desaturation and bradycardia events along with patient characteristics, including weight, sex, and gestational age [43]. This algorithm showed promising results, especially in a high negative predictive value (95%) but requires 72 hours of individual patient learning as well as combining vital signs with manually entered electronic health record data [52]. Such effectiveness could potentially allow the system to assist in decisions in antibiotic treatment to avoid unnecessary administration. A high NPV is important since a lot of infants receive antibiotics, possibly saving lives but also disturbing the microbiota. Antibiotics and deranged microbiota double the risk of NEC in preterm infants.

As several more continuous vital parameter assessments exists and seem to improve the prediction of life-threatening events, most systems currently under development use more parameters and include base characteristics of infants in the algorithms (Fig. 2). Human-in-the-loop is the

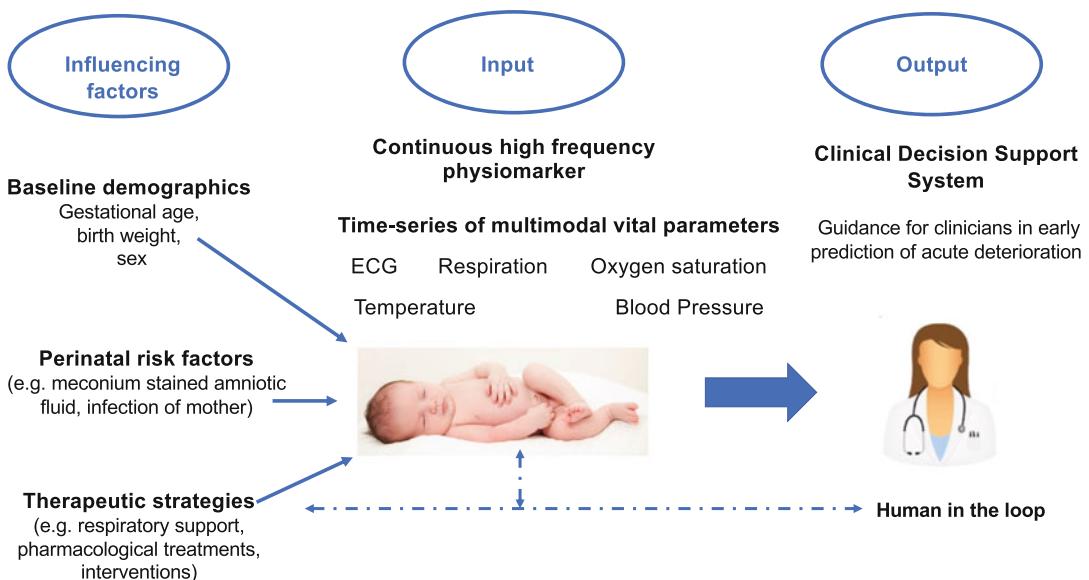


Fig. 2 Clinical Decision Support system - machine learning and human interaction Machine learning analysis of continuous multimodal complex and entangled physiological time series data –“Physiomarkers” is the basis for early detection of subtle alteration that help to predict adverse life-threatening events. The predictability

is improved by additional information. Both basic demographic data e.g. age, weight, sex and ongoing therapeutic interventions. Human in the loop machine learning-based trend analysis also involve tuning the model to improve its accuracy and guide adequate therapeutic strategies

process of leveraging the power of the machine and human intelligence to create ML-based AI models. Thus, humans annotate or label data then give to the ML algorithm to learn from such and take decisions from such predictions. And then humans also involve in tuning the model to improve its accuracy.

Pediatric Sepsis and Early Detection

Slightly older children are commonly monitored manually using bedside PEWS (Pediatric Early Warning Score), which relies on manually entered data by nurses at the patient bedside. This has been linked to a reduction in life-threatening events, but not deaths among hospitalized infants [40]. PEWS is widespread over the world but is labor-intensive and renders slight but limited benefits. Similar to the NICU, efforts of automatizing patient monitoring and developing biomarkers for sepsis have been made. Using logistic regression analysis, random forests, and convolutional neural network models with data on heart rate and blood pressure, patients with severe sepsis can be detected 8 h prior to fulfilling sepsis criteria [53]. Another random forest model was able to differentiate between sepsis and noninfectious systemic inflammatory response syndrome in PICU patients [54]. However, the model was based upon biomarkers in combination with other registered clinical events and thus did not utilize continuous automatic vital signs monitoring.

In children arriving at the pediatric emergency room (ER) with a diverting temperature (fever or hypothermia), ML models had a higher sensitivity in discriminating sepsis based upon vital signs compared to a routine physician judgment of the patient [55]. The ML models had however a lower specificity. An elastic net regularization model based upon patient demographic features in combination with blood pressure and body temperature measurements has also provided high negative predictive values for detecting septic shock in pediatric ER patients with a clinical suspicion of sepsis [56]. This model was not designed to screen all ER patients, but instead confirm suspicion of sepsis. Interestingly, vital

signs, such as heart rate, respiratory rate, and oxygen saturation, were evaluated but not included in the final model as they did not contribute to septic shock prediction. However, there is evidence that CDS systems developed for older children may be beneficial in ruling out sepsis [54–56], similar to the NICU systems.

Although not sepsis-specific, the Pediatric Risk of Mortality Prediction Tool (PROMPT) has been developed to estimate the mortality risk for PICU patients using vital sign monitoring. In a convolutional neural networks model, blood pressure, heart rate, respiratory rate, body temperature, and oxygen saturation were included as features. The prediction of death, independent of cause, was high as early as 60 h before the event.

Challenges and Future Perspectives of Automated Vital Signs Pattern Analysis

Although the automated vital sign pattern analyses for life-threatening event detection have shown great promise for clinical implementation, several challenges remain.

Important obstacles are the lack of consensus in outcome definition, heterogeneity in patient characteristics, and interpretability of data [57]. Can we trust in and interpret a deep learning-based ML-CDS? For the use of ML in the design of efficient and trusted CDS systems, it is important to develop an explainable ML (EML) for robust, reliable, and comprehensible data analysis. Explanations could help examine whether an ML method has employed true evidence instead of biases that widely exist in training data. This development of a new paradigm in EML, with focus on deep systems analysis of time series data, could be achieved by combining model-driven systems and data-driven learning methods. Models will help to design necessary constraints for analysis, in turn explainability. We hypothesize that data-driven optimization of proposed deep systems composed of many model-driven components will offer high quality performance along with explainability. Please see Yonina Eldar et al. chapter regarding mathematical foundations and explainable deep learning.

The majority of available pediatric CDS systems currently available that use continuous assessment of vital parameters are developed for the NICU and close to all based on preterm infants [13]. However, this is a heterogeneous population due to varying gestational ages at birth and birth weight, particularly when it comes to autonomic control [13, 25, 45, 58, 59].

The cardiorespiratory pattern changes over time and is dependent on gestational age, postnatal age, and postconceptual age [45, 58] and continue to develop during childhood. Most studies are conducted among neonates up to 1 week old, thus limiting its generalizability to older infants.

Sensors monitoring preterm infants' vital signs are often subject to movements. This creates an artifact on the signal, which may result in misleading or irrelevant features [60] attempted to clean the vital signs signal by recording movement around a patient. This enables discarding portions of vital signs disturbed by movements and may produce less noisy and more relevant signals. However and notably, infant movements might be used as a feature for detecting illness, as lethargy is a warning sign for sepsis [45, 61], and therefore discarding information may introduce a bias and result in suboptimal systems.

Including several parameters may improve early detection of life-threatening events [62]. In addition, more recently, there has been advances in video and audio processing, which provides the possibility of adding such features to algorithms [60, 63]. However, this requires additional hardware at the ICUs, potentially limiting system implementation. Using multiple features, and potentially a combination of features, may be the best way forward. Further research is required to identify the key ones. Pros and cons of different vital signs, and on which features these could be based have been discussed recently [13, 45].

Conclusions

Multiple attempts have been made to generate a functional CDS system for sepsis detection in young children. Although several of them have promising results, especially in the possibility of

excluding sepsis and other life-threatening events, very few have reached commercial availability. CDS systems generally allow detection of sepsis up to 24 h prior to current clinical praxis. Most systems are based on the analysis of heart rate variability and heart rate characteristics, but combining multiple features improves system performance. The use of CDS systems in study settings have been shown to allow earlier interventions and improve patient outcome. Currently, many systems are limited in that they are developed for narrow and specific patient subgroups, such as very low birth weight infants.

Every new system should be systematically evaluated in comparison to current systems already in use. International collaborations and data sharing would be fundamental in improving neonatal and pediatric care on a global scale, especially among intensive care unit patients. How to best approach ML model design, feature inclusion, and practical implementation are still issues that needs to be resolved. Research utilizing continuous real-time high-frequency vital sign monitor data to develop automated early warnings is crucial. Prospective evaluation using real-time data and decision-making in clinical trials and also testing in different healthcare settings is required, but a reduction in mortality from sepsis by only a small percentage would represent several tens of thousands of lives saved annually worldwide [64]. Therefore, further improving real-time algorithm-assisted clinical decision support systems to better guide treatments and improve outcomes, reduce morbidity and mortality of hospitalized infants is a crucial and exciting endeavor for the coming years.

References

1. Herlenius E, Kuhn P. Sudden unexpected postnatal collapse of newborn infants: a review of cases, definitions, risks, and preventive measures. *Transl Stroke Res.* 2013;4(2):236–47.
2. Fairchild K, Mohr M, Paget-Brown A, Tabacaru C, Lake D, Delos J, et al. Clinical associations of immature breathing in preterm infants: part 1 – central apnea. *Pediatr Res.* 2016;80(1):21–7.
3. Siljehav V, Hofstetter AM, Leifsdottir K, Herlenius E. Prostaglandin E2 mediates cardiorespiratory disturbances during infection in neonates. *J Pediatr.* 2015;167(6):1207–1213.e3.

4. Iroh Tam P-Y, Bendel CM. Diagnostics for neonatal sepsis: current approaches and future directions. *Pediatr Res.* 2017;82(4):574–83.
5. Adams-Chapman I. Long-term impact of infection on the preterm neonate. *Semin Perinatol.* 2012;36(6):462–70.
6. Patel RM, Kandefer S, Walsh MC, Bell EF, Carlo WA, Laptook AR, et al. Causes and timing of death in extremely premature infants from 2000 through 2011. *N Engl J Med.* 2015;372(4):331–40.
7. Schindler T, Koller-Smith L, Lui K, Bajuk B, Bolisetty S, New South Wales and Australian Capital Territory Neonatal Intensive Care Units' Data Collection. Causes of death in very preterm infants cared for in neonatal intensive care units: a population-based retrospective cohort study. *BMC Pediatr.* 2017;17(1):59.
8. Cailes B, Kortsalioudaki C, Buttery J, Pattayak S, Greenough A, Matthes J, et al. Epidemiology of UK neonatal infections: the neonIN infection surveillance network. *Arch Dis Child Fetal Neonatal Ed.* 2018;103(6):F547–53.
9. Fleischmann-Struzek C, Goldfarb DM, Schlattmann P, Schlapbach LJ, Reinhart K, Kissoon N. The global burden of paediatric and neonatal sepsis: a systematic review. *Lancet Respir Med.* 2018;6(3):223–30.
10. Rudd KE, Johnson SC, Agesa KM, Shackelford KA, Tsoi D, Kievlan DR, et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. *Lancet.* 2020;395(10219):200–11.
11. Ohlin A, Björkqvist M, Montgomery SM, Schollin J. Clinical signs and CRP values associated with blood culture results in neonates evaluated for suspected sepsis. *Acta Paediatr Oslo Nor* 1992. 2010;99(11):1635–40.
12. Shane AL, Sánchez PJ, Stoll BJ. Neonatal sepsis. *Lancet Lond Engl.* 2017;390(10104):1770–80.
13. Kumar N, Akangire G, Sullivan B, Fairchild K, Sampath V. Continuous vital sign analysis for predicting and preventing neonatal diseases in the twenty-first century: big data to the forefront. *Pediatr Res.* 2020;87(2):210–20.
14. Cortese F, Scicchitano P, Gesualdo M, Filaninno A, De Giorgi E, Schettini F, et al. Early and late infections in newborns: where do we stand? A review. *Pediatr Neonatol.* 2016;57(4):265–73.
15. Singer M, Deutszman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA.* 2016;315(8):801.
16. Schlapbach LJ, Kissoon N. Defining pediatric sepsis. *JAMA Pediatr.* 2018;172(4):313.
17. Wynn JL, Wong HR, Shanley TP, Bizzarro MJ, Saiman L, Polin RA. Time for a neonatal-specific consensus definition for sepsis. *Pediatr Crit Care Med.* 2014;15(6):523–8.
18. Menon K, Schlapbach LJ, Akech S, Argent A, Chiotos K, Chisti MJ, et al. Pediatric sepsis definition – a systematic review protocol by the pediatric sepsis definition taskforce. *Crit Care Explor.* 2020;2(6):e0123.
19. Kuhn P, Messer J, Paupe A, Espagne S, Kacet N, Mouchino G, et al. A multicenter, randomized, placebo-controlled trial of prophylactic recombinant granulocyte-colony stimulating factor in preterm neonates with neutropenia. *J Pediatr.* 2009;155(3):324–330.e1.
20. Auriti C, Fiscarelli E, Ronchetti MP, Argentieri M, Marrocco G, Quondamcarlo A, et al. Procalcitonin in detecting neonatal nosocomial sepsis. *Arch Dis Child Fetal Neonatal Ed.* 2012;97(5):F368–70.
21. Wu T-W, Tabangin M, Kusano R, Ma Y, Ridsdale R, Akinbi H. The utility of serum hepcidin as a biomarker for late-onset neonatal sepsis. *J Pediatr.* 2013;162(1):67–71.
22. Wynn JL. Defining neonatal sepsis. *Curr Opin Pediatr.* 2016;28(2):135–40.
23. Gordon A, Jeffery HE. Antibiotic regimens for suspected late onset sepsis in newborn infants. *Cochrane Database Syst Rev.* 2005;3:CD004501.
24. Marik PE, Taeb AM. SIRS, qSOFA and new sepsis definition. *J Thorac Dis [Internet].* 2017 Apr [cited 2021 Mar 31];9(4). <https://jtd.amegroups.com/article/view/12738>
25. Herlenius E. An inflammatory pathway to apnea and autonomic dysregulation. *Respir Physiol Neurobiol.* 2011;178(3):449–57.
26. Gang Y, Malik M. Heart rate variability in critical care medicine. *Curr Opin Crit Care.* 2002;8(5):371–5.
27. Mangoni ME, Nargeot J. Genesis and regulation of the heart automaticity. *Physiol Rev.* 2008;88(3):919–82.
28. Bainton CR, Richter DW, Seller H, Ballantyne D, Klein JP. Respiratory modulation of sympathetic activity. *J Auton Nerv Syst.* 1985;12(1):77–90.
29. Trowbridge HO. Inflammation: a review of the process. 5th ed. Chicago: Quintessence; 1997.
30. Ek M, Engblom D, Saha S, Blomqvist A, Jakobsson PJ, Ericsson-Dahlstrand A. Inflammatory response: pathway across the blood-brain barrier. *Nature.* 2001;410(6827):430–1.
31. Engblom D, Saha S, Engstrom L, Westman M, Audoly LP, Jakobsson PJ, et al. Microsomal prostaglandin E synthase-1 is the central switch during immune-induced pyrexia. *Nat Neurosci.* 2003;6(11):1137–8.
32. Hofstetter AO, Saha S, Siljehav V, Jakobsson P-J, Herlenius E. The induced prostaglandin E2 pathway is a key regulator of the respiratory response to infection and hypoxia in neonates. *Proc Natl Acad Sci.* 2007;104(23):9894–9.
33. Kim SO, Dozier BL, Kerry JA, Duffy DM. EP3 receptor isoforms are differentially expressed in subpopulations of primate granulosa cells and couple to unique G-proteins. *Reproduction.* 2013;146(6):625–35.
34. Kawahara K, Hohjoh H, Inazumi T, Tuchiya S, Sugimoto Y. Prostaglandin E2-induced inflammation: relevance of prostaglandin E receptors. *Biochim Biophys Acta BBA – Mol Cell Biol Lipids.* 2015;1851(4):414–21.

35. Forsberg D, Horn Z, Tsgera E, Smedler E, Silberberg G, Shvarev Y, et al. CO₂-evoked release of PGE2 modulates sighs and inspiration as demonstrated in brainstem organotypic culture. *elife.* 2016;5:e14170.
36. Guerra FA, Savich RD, Wallen LD, Lee CH, Clyman RI, Mauray FE, et al. Prostaglandin E2 causes hypoventilation and apnea in newborn lambs. *J Appl Physiol.* 1988;64(5):2160–6.
37. Hofstetter A, Legnevall L, Herlenius E, Katz-Salamon M. Cardiorespiratory development in extremely preterm infants: vulnerability to infection and persistence of events beyond term-equivalent age: cardiorespiratory development in extremely preterm infants. *Acta Paediatr.* 2008;97(3):285–92.
38. Siljehav V, Olsson Hofstetter A, Jakobsson P-J, Herlenius E. mPGES-1 and prostaglandin E2: vital role in inflammation, hypoxic response and survival. *Pediatr Res.* 2012;72(5):460–7.
39. Siljehav V, Shvarev Y, Herlenius E. IL-1 β and prostaglandin E₂ attenuate the hypercapnic as well as the hypoxic respiratory response via prostaglandin E receptor type 3 in neonatal mice. *J Appl Physiol.* 2014;117(9):1027–36.
40. Parshuram CS, Dryden-Palmer K, Farrell C, Gottesman R, Gray M, Hutchison JS, et al. Effect of a pediatric early warning system on all-cause mortality in hospitalized pediatric patients: the EPOCH randomized clinical trial. *JAMA.* 2018;319(10):1002–12.
41. Kowalski RL, Lee L, Spaeder MC, Moorman JR, Keim-Malpass J. Accuracy and monitoring of pediatric early warning score (PEWS) scores prior to emergent pediatric intensive care unit (ICU) transfer: retrospective analysis. *JMIR Pediatr Parent.* 2021;4(1):e25991.
42. Sullivan BA, McClure C, Hicks J, Lake DE, Moorman JR, Fairchild KD. Early heart rate characteristics predict death and morbidities in preterm infants. *J Pediatr.* 2016;174:57–62.
43. Mithal LB, Yoge R, Palac HL, Kaminsky D, Gur I, Mestan KK. Vital signs analysis algorithm detects inflammatory response in premature infants with late onset sepsis and necrotizing enterocolitis. *Early Hum Dev.* 2018;117:83–9.
44. Honoré A, Liu D, Forsberg D, Coste K, Herlenius E, Chatterjee S, et al. Hidden Markov Models for sepsis detection in preterm infants. In: ICASSP 2020–2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). 2020. p. 1130–4.
45. Joshi R, Kommers D, Oosterwijk L, Feij L, van Pul C, Andriessen P. Predicting neonatal sepsis using features of heart rate variability, respiratory characteristics, and ECG-derived estimates of infant motion. *IEEE J Biomed Health Inform.* 2020;24(3):681–92.
46. Moss TJ, Lake DE, Calland JF, Enfield KB, Delos JB, Fairchild KD, et al. Signatures of subacute potentially catastrophic illness in the ICU: model development and validation*. *Crit Care Med.* 2016;44(9):1639–48.
47. Hicks JH, Fairchild KD. Heart rate characteristics in the NICU: what nurses need to know. *Adv Neonatal Care.* 2013;13(6):396–401.
48. Fairchild K, Aschner J. HeRO monitoring to reduce mortality in NICU patients. *Res Rep Neonatol.* 2012;2012(default):65–76.
49. Sullivan BA, Fairchild KD. Predictive monitoring for sepsis and necrotizing enterocolitis to prevent shock. *Semin Fetal Neonatal Med.* 2015;20(4):255–61.
50. Moorman JR, Carlo WA, Kattwinkel J, Schelonka RL, Porcelli PJ, Navarrete CT, et al. Mortality reduction by heart rate characteristic monitoring in very low birth weight neonates: a randomized trial. *J Pediatr.* 2011;159(6):900–906.e1.
51. Sullivan B, Wallman-Stokes A, Isler J, Sahni R, Moorman J, Fairchild K, et al. Early pulse oximetry data improves prediction of death and adverse outcomes in a two-center cohort of very low birth weight infants. *Am J Perinatol.* 2018;35(13):1331–8.
52. Gur I, Riskin A, Markel G, Bader D, Nave Y, Barzilay B, et al. Pilot study of a new mathematical algorithm for early detection of late-onset sepsis in very low-birth-weight infants. *Am J Perinatol.* 2014;32(04):321–30.
53. Kamaleswaran R, Akbilgic O, Hallman MA, West AN, Davis RL, Shah SH. Applying artificial intelligence to identify physiomarkers predicting severe sepsis in the PICU. *Pediatr Crit Care Med.* 2018;19(10):e495–503.
54. Lampert F, Jack T, Rübsamen N, Sasse M, Beerbaum P, Mikolajczyk RT, et al. Development and validation of a diagnostic model for early differentiation of sepsis and non-infectious SIRS in critically ill children – a data-driven approach using machine-learning algorithms. *BMC Pediatr.* 2018;18(1):112.
55. Balamuth F, Alpern ER, Grundmeier RW, Chilutti M, Weiss SL, Fitzgerald JC, et al. Comparison of two sepsis recognition methods in a pediatric emergency department. *Acad Emerg Med.* 2015;22(11):1298–306.
56. Scott HF, Colborn KL, Sevick CJ, Bajaj L, Kissoon N, Deakyne Davies SJ, et al. Development and validation of a predictive model of the risk of pediatric septic shock using data known at the time of hospital arrival. *J Pediatr.* 2020;217:145–151.e6.
57. Monga V, Li Y, Eldar YC. Algorithm unrolling: interpretable, efficient deep learning for signal and image processing. *IEEE Signal Process Mag.* 2021;38(2):18–44.
58. Javorka K, Lehotska Z, Kozar M, Uhríková Z, Kolarovszki B, Javorka M, et al. Heart rate variability in newborns. *Physiol Res.* 2017;66(Suppl 2):S203–14.
59. Stojanovska V, Miller SL, Hooper SB, Polglase GR. The consequences of preterm birth and chorioamnionitis on brainstem respiratory centers: implications for neurochemical development and altered functions by inflammation and prostaglandins. *Front Cell Neurosci.* 2018;12:26.
60. Värrö A, Kallonen A, Helander E, Ledesma A, Pladys P. The Digi-NewB project for preterm infant sepsis risk and maturity analysis. *Finn J EHealth EWelfare [Internet].* 2018 May 21 [cited 2019 Jun 4];10(2–3). <https://journal.fi/fnjehew/article/view/69152>

61. Verstraete EH, Blot K, Mahieu L, Vogelaers D, Blot S. Prediction models for neonatal health care-associated sepsis: a meta-analysis. *Pediatrics*. 2015;135(4):e1002–14.
62. Griffin MP, Lake DE, Bissonette EA, Harrel FE, O’Shea TM, Moorman JR. Heart rate characteristics: novel biomarkers to predict neonatal infection and death. *Pediatrics*. 2005;116(5):1070–4.
63. Cabon S, Porée F, Simon A, Rosec O, Pladys P, Carrault G. Video and audio processing in paediatrics: a review. *Physiol Meas*. 2019;40(2):02TR02.
64. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med*. 2018;24(11):1716–20.



Aging and Alzheimer's Disease

76

Application of Artificial Intelligence in Mechanistic Studies, Diagnosis, and Drug Development

Ruixue Ai, Xurui Jin, Bowen Tang, Guang Yang,
Zhangming Niu, and Evandro F. Fang

Contents

Introduction	1058
AI in Healthcare	1059
Historical Overview	1059
AI for Drug Discovery	1060
AI in Biology	1062
AI in Medicine	1063
Application of Machine Learning in Clinical Work for Alzheimer's Disease ...	1065
Etiology	1065

R. Ai

Department of Clinical Molecular Biology, University of
Oslo, Oslo, Norway

Akershus University Hospital, Lørenskog, Norway

X. Jin · B. Tang

MindRank AI Ltd., Hangzhou, Zhejiang, China

G. Yang

Cardiovascular Research Centre, Royal Brompton
Hospital, London, UK

National Heart and Lung Institute, Imperial College

London, London, UK

e-mail: g.yang@imperial.ac.uk

Z. Niu

MindRank AI Ltd., Hangzhou, Zhejiang, China

Aladdin Healthcare Technologies Ltd., London, UK

e-mail: zhangming@mindranks.ai

E. F. Fang (✉)

Department of Clinical Molecular Biology, University of
Oslo and Akershus University Hospital, Lørenskog,
Norway

The Norwegian Centre on Healthy Ageing (NO-Age),

Oslo, Norway

e-mail: e.f.fang@medisin.uio.no

Diagnosis	1066
Therapy	1066
Prognosis	1068
Future Perspectives and Concluding Remarks	1068
References	1069

Abstract

Artificial intelligence (AI) implies the use of a machine with limited human interference to model intelligent actions. It covers a broad range of research studies from machine intelligence for computer vision, robotics, and natural language processing to more theoretical machine learning algorithm design and, recently, “deep learning” development. The application of AI in medical fields is booming, including the use of AI in data collection, analysis, mechanistic prediction, to clinical disease diagnosis and drug development. In this chapter, we focus on the challenges in the studies of aging and the age-predisposed Alzheimer’s disease (AD) and summarize on how to use AI to help addressing these questions. We finally provide future perspectives on the use of AI in aging research and AD.

Introduction

Artificial intelligence (AI) is a generic concept that implies the use of a machine with limited human interference to model intelligent actions. It covers a broad range of research studies from machine intelligence for computer vision, robotics, and natural language processing to more theoretical machine learning algorithm design and, recently, “deep learning” development. In general, AI is recognized as having begun with the invention of robotics in the 1920s. There have been several waves of success and stagnancy over the years, most recently exemplified in the recent breakthrough powered by the development of more powerful graphics processing units (GPUs) and the outbreak of big data. Funding for both AI-based research studies and industrial innovation projects has further propelled progress and accelerated development.

In medicine, artificial intelligence (AI) research is growing rapidly with broad applications (Fig. 1). In 2016, healthcare AI projects attracted more investment in comparison with AI projects from other sectors of the global economy. In addition to conventional mathematical and statistical methods, AI techniques, in particular machine learning and deep learning approaches, draw significant attention to the analysis of medical data as medicine is becoming an increasingly data-centric discipline. The nature of evidence-based medicine is to guide therapeutic decision-making through learning from past data. Statistical approaches have historically addressed this challenge by characterizing correlations inside data through statistical equations, such as linear regression, indicating a “line of best fit.” Through machine learning or deep learning, AI can reveal nuanced relationships, which cannot simply be deduced by an equation. AI can extract valuable details from the electronic footprint of a patient. This will initially save time and increase performance but can also explicitly guide patient management after appropriate research. Specialist diagnostic skills can now also be carried into primary care by AI-based technologies. Broadly speaking, there are three types of algorithms for machine learning including (1) unsupervised learning (capable of identifying patterns directly from data, e.g., activity recognition using smartphone sensors, clinical outcome prediction from the electronic health records, anomaly detection, and knowledge discovery), (2) (semi-)supervised learning (classification and inferencing based on previous examples, e.g., chemoinformatics and drug discovery, multifactorial omics analysis and biomarker identification, imaging, and data enhancement), and (3) reinforcement learning (synergy of reward and punishment sequences to form an operating strategy in a given problem space, e.g., therapy optimization, treatment management, and surgical robotics).

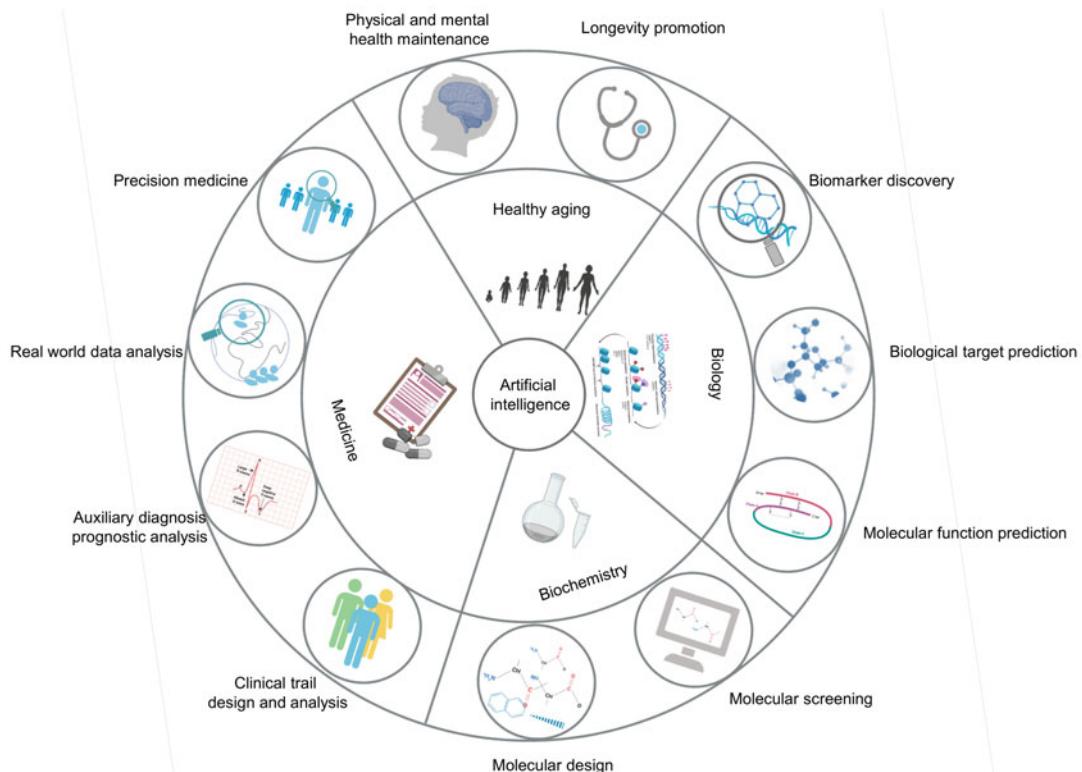


Fig. 1 The use of artificial intelligence (AI) in healthcare. The broad application of AI in healthcare consists of overlapping fields like biology, biochemistry, medicine, and healthy aging. For details see the text

In general, AI techniques have returned significant dividends by achieving outstanding performance in various applications especially for AI in medicine. There are a few key problems that need to be solved. For example, most current AI-based algorithms are typical black boxes. The end users can easily access high-performance AI, but the interpretability of the derived features and prediction is not clear. Therefore, explainable AI will be a major area for future research. Moreover, although current studies rely on a human operator's manual labelling, the ultimate goal will be "Taking the Human Out of the Loop." This encompasses future research directions including self-supervised learning, semi-supervised learning, and more advanced non-supervised learning methods. The first sections of this chapter will detail the historical development of AI in healthcare and in particular AI in biology, biochemistry, medicine, and healthy aging. The

rest of the chapter will focus on the application of AI in clinical work for Alzheimer's disease including diagnosis, treatment planning, drug development, and prognosis. We will also shed a light on future perspectives.

AI in Healthcare

Historical Overview

Among the broad range of research studies covered by machine intelligence, medicine has been identified as one of the most important areas for AI applications. Since the advent of AI, applications of AI in medicine have been explosive and make it possible for more precise and personalized medicine. The progressive growth and development of the application of AI in medicine can be organized by specific periods.

From the 1950s to the 2000s

In the first 20 years of this stage, researchers in this field primarily focused on the development of decision-making systems that had the ability to make inferences or decisions that could be applied in the clinical setting. In this period, some clinical informatics databases and medical record systems were firstly built, which had established the foundation for future applications of AI in medicine [1]. The last 30 years of this period was referred to as the AI winter. During this time, research funding and interest in AI in general were significantly reduced, and accordingly there was slow progress in the field. Generally, two major historical limitations had led to this famous “winter” which was also reflected in its application in medicine. The first was the extensive cost of building and maintaining the expert-labelled databases and obtaining the necessary calculating power. The second was the perceived limitations of AI such as the efficiency of the algorithm. Although there was a lack of funding and interest from academia during this period, collaboration among pioneers in the field of AI and medicine continued. In 1973, the Stanford University Medical Experimental Computer-Artificial Intelligence in Medicine (SUMEX-AIM) project was founded to enhance the connectivity between computer scientists and biomedical researchers from several institutions [1]. In 1975, the NIH sponsored an AI in medicine workshop which was held in Rutgers University [2]. Those initial collaborations among the pioneers in the field facilitated AI’s application to medicine at this stage.

In later years of this period (after the 1970s), research turned toward comparing the performance of computer-aided diagnostic systems with that of human physicians [3]. Some systems had also been shown to have the ability to interpret ECGs, diagnose diseases, choose appropriate treatments, provide interpretations of clinical reasoning, and assist physicians in generating diagnostic hypotheses in complex patient cases [4–6]. Scientists also put some research effort into the use of a computer to improve the detection and classification of the lesions in the vascular and integumentary systems using image analysis.

After the 2000s

As the development of computer hardware accelerated, the application of AI in medicine started to grow fast. Although the development of neural networks had begun in 1950, by the 1980s, it had progressed to such a point that it was now possible to create deep learning models (a phrase first coined by Rina Dechter in 1986) [7]. The development of deep learning (DL) marked important progress. Although deep neural networks had technically first been studied in the 1950s, their application to medicine had been limited by two problems for many years: first, insufficient computing capacity and second, the lack of available training data [8]. These limitations were overcome in the 2000s with increases in the availability of large data sources and significantly improved computing power. The large data sources included large-scale studies (e.g., UK Biobank), data-collection platforms (e.g., Broad Bioimage Benchmark Collection and the Image Data Resource), and electronic medical record (EMR) data [9, 10]. To highlight this increase, a national survey conducted in 2008 noted that only 13% of physicians reported having a basic EMR system, but by the end of 2012, 72% of physicians had adopted some type of EMR system [11, 12]. Since that time, the application of AI in medicine has been on the “fast track.” With the advances came substantial progress in image classification tasks and innovations in the application of AI to other fields such as drug development.

AI for Drug Discovery

The discovery of molecular structures with specifically required properties has been one of the most impactful scientific and industrial challenges. Normally, to develop a novel drug, the mean pre-tax expenditure is nearly \$2.55 billion over 10–15 years [13]. In recent years as more and more big chemical datasets have become available, deep learning has been applied to those data to accelerate the process of drug discovery and reduce costs in the process. AI technology can be applied during several steps of the drug design process, including virtual screening, activity

scoring, quantitative structure-activity relationship (QSAR) analysis, de novo drug design, and in silico evaluation of absorption, distribution, metabolism, excretion, and toxicity (ADME/T) properties [14].

Virtual Screening

Virtual screening (VS) is defined as the use of algorithms to find bioactive molecules from known molecular libraries for a drug target, typically an enzyme or a protein receptor. It has proved to be a very effective approach to filtering out those compounds with unfavorable properties [15, 16]. Based on traditional computer-aided drug discovery (CADD), virtual screening usually uses 2D and 3D structural information from ligands and target proteins. Similarity searching of small molecules, pharmacophore-matching, and molecular docking are commonly used techniques in traditional virtual screening. All the abovementioned techniques depend on the knowledge of chemistry, physics, and biology. While the technology used to run these CADD methods for virtual screening is affordable and easy to learn, it is hard to master them to successfully find the bioactive molecule for a given target. Recently, deep learning has been applied to virtual screening [17]. As opposed to traditional knowledge-based methods, deep learning VS directly extrapolates the rules through examination of its training data. Using molecular features, it tries to predict the classes of bioactivity through the examination of many abstract and high-level features that are sometimes not immediately obvious to the human eye. Due to the high feature extraction ability and low generalization error of deep learning, it is especially well adapted for use in ligand-based virtual screening which does not rely on the 2D/3D structural information of proteins [18–21]. For example, traditionally, the sparse distribution of active compounds in the general database wasted a lot of time during the process of virtual screening [22]. DL methods have been used to address this issue through use of the long short-term memory (LSTM) network model. The model was based on the similarity between natural language and simplified molecular-input line-entry system (SMILES) which is a specification in the form of a line notation for describing the

structure of chemical species using short ASCII strings [23]. Using AI for virtual screening may increase the chance of identifying new targets and make the process more rational. Another implementation of AI virtual screening is the use of abundant data from high-throughput experiments with gene expression profiles. For example, researchers tried to find potential drugs through measurement of the functional similarity of small molecules based on gene expression data [24].

Bioactivity Scoring

Activity scoring is one of the core components of molecular docking, a process that evaluates the potential binding affinities of drug-like molecules toward an interested target [25]. The DL-based methods perform well in this field due to their high nonlinear mapping ability and the fact that they could extract features efficiently from genomic, chemical, and physical force data [26]. For instance, some studies using convolutional neural networks (CNNs) to extract the features from protein-ligand interaction images were able to predict protein-ligand affinity [27]. A study using a 3D CNN model demonstrated prediction of binding affinities that was well matched with the experimental data [27]. Additionally, some studies used DLs to extract features from basic and primitive features. DeepVS was built with a CNN model and was designed to discern abstract features from basic chemical features (e.g., the atom context). DeepVS outperformed the traditional docking programs on area under the curve (AUC) and enrichment factor [28].

ADME/T Properties Prediction

It is vital to identify the molecules with poor chemical properties in the drug discovery pipeline. Early identification of ADME/T properties can reduce the risk of failure and save a large amount of time and money during development. Many DL-based methods were developed to address this issue [29]. One study utilized a CNN-ANN that extracted data from the molecular graph, and another study used the tensor-based convolutional embedding of attributed molecular graphs method to predict the solubility of molecules [30, 31]. The two models both showed good predictive performance

in their testing data. Some studies that focused on predicting drug absorption, the process by which drugs entered the blood from the site of administration, also applied DL-based methods. For example, a study with 1,014 molecules used the MLR model to predict bioavailability with structural fingerprints and molecular properties [32]. The model had a good predictive performance (correlation coefficient 0.71, MSE 0.24). The DL-based methods have also improved predictive performance in modelling drug distribution, metabolism, excretion, and toxicity. Recently, some multitask DL models for ADME/T prediction were also developed and showed good performance compared with other models targeted on predicting single property [33].

De Novo Drug Design

The de novo molecule generation problem involves generating novel or modified molecular structures with desirable properties. Thus, generative AI models are usually coupled with the abovementioned predictive neural networks to generate new compound structures under the constraints of interesting molecular properties. For example, Mariya et al. proposed ReLeASE which combined a molecular property predicting neural network with a molecular generative neural network to design chemical compounds with desired physical, chemical, and/or bioactivity properties [34]. However, Mariya et al. did not combine their generative AI model with bioactive experiments in their research. In silico medicine built the generative AI model GENTRL to auto-design novel inhibitors for kinase DDR1; their generated chemical compounds were identified as bioactive by their wet labs [35]. Additionally, because of the high generative ability of DL models, a study used data from the NCI-60 cell line assay to train an adversarial autoencoder model. This model could be used to generate the molecular fingerprints that were helpful in the search for potential anticancer agents.

AI in Biology

DL can be applied in the field of biology to answer some fundamental questions because it is suitable

for dealing with high-dimension data from omics, such as genomics and proteomics.

Genetics

Currently, a massive amount of genomics data is produced using next-generation sequencing technology, and AI is applied in analyzing those data. There are an increasing number of studies using DL to examine the genome and functional genomics. For example, DL has been applied in: (1) predicting the sequence specificity of DNA- and RNA-binding proteins, (2) methylation status, (3) gene expression, and (4) control of splicing [36]. Additionally, DL has been successfully applied in regulatory genomics. In this field, some architectures from computer vision or natural language processing were well suited after some genomic-specific modifications [36].

DL can also answer the question “how much RNA is produced from a DNA template in a particular cell” by building a model to predict gene expression from genotype data. It can be used for studying splicing-code models as well as for the identification of long noncoding RNAs. DL has been used for the interpretation of regulatory control in single cells [37], for example, the detection of DNA methylation in single cells, and for the identification of subgroups of cells through improvement of the representation of single-cell RNA-seq data. Additionally, predicting phenotypes is also one of the major interests of DL in genomics [36].

Proteomics

Proteomics aims to study the proteins’ structure and function in biological systems and is gradually becoming a data-rich discipline. Accordingly, DL is warranted to interpret the huge amount of data, giving biological insights. In the field of proteomics, DL performed well in predicting protein structure, posttranslational modification, and MHC-binding peptide [38].

The function of a protein largely depends on its structure, and predicting the spatial structure from an amino acid sequence plays a vital role in protein design and drug screening [39]. Currently, nearly 170,000 protein structures have been measured through a multicenter effort; however, it is a very

time-consuming and high-cost process to directly measure the structure. Using DL to predict the structure may reduce the cost and accelerate the related research [38]. Recently, the performance protein structural predictive models was significantly improved through the application of AI, especially in those predictions without previously known homologous structures. After 11 rounds of Critical Assessment of protein Structure Prediction (CASP), the performance of DL in predicting protein structures increased quickly, especially in terms of development of residue-residue contact predictions [40]. In the most recent CASP14, the performance of AlphaFold2 achieved remarkably high improvement, and its success rate received considerable interest from both academia and the general public. Although the structure of AlphaFold2 has not been published yet, we know that its precursor AlphaFold has a highly complex, dilated residual neural network (ResNet) with 220 blocks to predict the C β distances of residue pairs given the amino acid sequence and many MSA-derived features [41].

AI in Medicine

Diagnosis

Image-based diagnosis has been regarded as the most successful application of AI in medicine. Its application scenarios in the hospital are widely used in radiology, ophthalmology, dermatology, and pathology to assist with image-based diagnoses [42].

In the department of radiology, the earliest application of computer-assisted diagnosis may date back to the 1970s [43]. Currently, with the development of methodology, AI was applied in the detection of lung nodules and the diagnosis of pulmonary tuberculosis and other common lung diseases using images from chest radiography [44, 45]. Additionally, breast-mass identification using mammography scans reached expert-level diagnostic accuracies [46, 47]. Currently, many clinical diagnostic AIs are seeking legal approval for clinical applications. For example, an AI system for cardiovascular disease diagnosis with MRI image was registered by the FDA in 2018.

In dermatology, physicians use AI to diagnose various skin lesions and further differentiate disease. A recent study suggested that in diagnosing skin malignancy, convolutional neural networks achieved dermatologist-level accuracy [48]. The DL model performed better than the dermatologist in a comparison of algorithm predictions to the assessments of 21 dermatologists given a set of photographic and dermoscopic images. Although the training phase of the deep learning model can be expensive and time-consuming, the final model can easily be used on mobile devices which is very convenient and fast for disease screening [48].

Fundus photography is a noninvasive procedure that uses retinal cameras to capture images of the retina, optic disc, and macula [49]. Images acquired by fundus photography can be used as another source of data for the AI-assisted diagnosis. Recently, a research team of computer scientists and clinicians trained convolutional neural network models to identify referable diabetic retinopathy and diabetic macular edema with 128,175 retinal images. In this study, the DL model had a good performance in two independent datasets (areas under the receiver operating characteristic curve > 0.99) [50]. Additionally, its performance was comparable to the performance of expert-level ophthalmologists. This study also demonstrated that DL could identify the underlying associations between the images and age, gender, systolic blood pressure, smoking status, or cardiovascular events, which indicated the ability of DL to elicit new knowledge from raw data [51]. Another study showed that the performance of a convolutional neural network exceeded in pre-specified sensitivity (85%) and specificity (82.5%); it was authorized by the FDA for use by healthcare providers to detect diabetic macular edema and moderate-to-severe diabetic retinopathy [52].

In the Department of Pathology, images from histopathological assessment are used with AI algorithms to assist diagnosis. AI can be applied in the detection of various cancers and their metastases using biopsy specimens. For instance, DL models used in conjunction with normal clinical assessments can facilitate risk stratification of prostate and breast cancer patients [53]. In the USA, it is estimated that there will be a deficit of

more than 5,700 pathologists in the next 10 years [54]. DL detection system may be able to mitigate this gap and provide a fast and accurate assessment from histopathological slides or other biopsy specimens, further improving the quality of care for cancer patients.

Overall, successful applications of AI in radiology, dermatology, ophthalmology, and pathology have been followed by the availability of large labelled datasets, improved computational power, and the development of deep learning methods. Currently, those applications are changing medical practice dependably.

Prognosis

As more and more longitudinal data and EHRs become available, the DL model can learn from the trajectories of a large number of patients and further predict their prognosis. For instance, whether the patient could go back to work in a short period or how long they may have before the disease progress could be predicted. Several large integrated health systems included a DL model to evaluate the risk of transferring to the intensive care unit for in-hospital patients [55]. Additionally, some studies built DL models to predict mortality, readmission, and length of hospital stay using data from EHRs, classifying the cancer patients with different responses to chemotherapy. At the large population level, the same type of forecasting could perform risk stratification and identify those patients who have a higher risk of readmission or may need more healthcare services. Such information could lead to a better healthcare resource allocation and provide evidence-based healthcare [56].

One limitation for such models is the integration of the health data. Building prognosis systems needs longitudinal data to provide a comprehensive view of the patients' disease course. The integrated data need to include outcomes such as mortality, readmission, and medical cost; however, those data are held by a variety of bodies such as hospitals, the local public health bodies, medical insurance departments, and public security bureau. To integrate those data from different sources, a better solution could be to put those data in the hands of patients through ID

cards or other personal terminals. Additionally, standardization of the collected data is a vital issue for building such models [56].

Aging Biomarker Development

Aging is a gradual, multi-organ process that leads to multiple age-related diseases. Currently, some new experimental techniques have produced a huge amount of aging-related high-dimensional data. These data can evaluate the aging process precisely. AI technologies are suitable for the identification of the biomarkers of aging through patterns within these high-dimensional datasets [35].

Some DL-based aging clocks have already been created using data from genomics, biochemistry, proteomics, and clinical imaging. Biological age is a more precise measure of aging compared to chronological age. Generally, changes in biological age can provide a more comprehensive and objective evaluation of general health status [57, 58]. Data from MRI has been used for the determination of biological age, which can serve as a biomarker of aging. A study using a deep neural network (DNN)-based DL method on T1-weighted MRI images was shown to outperform random forest- or ANN-based age prediction models, with the predicted age named the brain age gap (BAG) [59]. The results demonstrated that age could be accurately predicted using unimodal imaging in a young population using engineered features instead of raw images. Notably, the discrepancy between the BAG and chronological age could be regarded as a marker of aging speed, and BAG could be used as a biomarker for neurodevelopment and disease detection that could be easily interpreted. Additionally, the biological age also can be built with transcriptomic data using deep learning. Some studies built deep learning models using transcriptomic data were presented. In one study using data from 545 transcriptomic samples from 12 datasets of human skeletal muscle, a deep feature selection (DFS) model was built and compared with several regression models. Linear regression was used as a baseline, and its performance was compared with other DL approaches. Although all models achieved a strong correlation

between predicted and chronological age, SVM and DFS models clearly outperformed the other methods in age prediction (R^2 0.83/0.83, mean absolute error 7.20/6.24 years). Some studies also used clinical EMR data to estimate biological age using DL. One study used DNNs to predict human chronological age using 41 biomarkers extracted from thousands of blood biochemistry samples from patients undergoing routine physical examinations. The DNN model with the best performance had an R^2 of 0.80 with an MAE of 6.07 years. In the analysis of feature importance, albumin, glucose, alkaline phosphatase, erythrocytes, and urea were identified as the five most important variables. These five biomarkers could also be identified from within thousands of biomarkers by DL [35].

Application of Machine Learning in Clinical Work for Alzheimer's Disease

Etiology

Alzheimer's disease (AD) is a neurodegenerative disease. Patients experience a steady progression of dementia, finally losing the ability to respond appropriately to their environment.

Currently, the pathogenesis of AD is still unclear although it is thought to be caused by an interaction between genetic and environmental factors [60, 61], with genetic factors contributing approximately 70% [62–65]. One of the prevalent theories of AD pathogenesis is the amyloid hypothesis which holds that different factors cause an imbalance between β -amyloid production and clearance which leads to β -amyloid accumulation in the brain. In turn, this accumulation leads to the formation of neurofibrillary tangles and neuroinflammation. As a result, the neurons may eventually become dysfunctional and die. Recently, the mitophagy pathway has also been found to contribute to AD [66], but the detailed pathology is still unclear [62].

Etiological studies are aiming to discover the environmental and genetic factors causing disease, and these can be clues for researching both

the prevention and treatment of AD. Recently, the development of AI technology has made it possible to achieve this goal using big data and ultra-complex models that exceed human brain processing capabilities [67–69]. AI provides a new way to model neuronal biological components using pathophysiological functional modules that can be embedded to model the complex dynamics influencing neuropsychiatric disorder phenomenology [70]. Because genetic factors contribute to approximately 70% of AD cases, they have been the main focus of AD pathogenesis research. In recent years, wide use of microarray and next-generation sequencing technologies has allowed research using genetic data to grow explosively. AI technology is becoming urgently required. Currently, genetic research on AD with AI is continuously growing.

In the population, individual genetic variations include (1) aneuploidy or polyploidy; (2) chromosome rearrangements including duplication, deletion, and inversion as well as translocations; (3) large segment deletions and duplications; (4) small insertions and deletions; (5) tandem repeat variations; and (6) single nucleotide variations (SNVs) [71]. There are approximately 3.2×10^9 base pairs (bp) in the genome of human beings, but 99% of those form noncoding regions. These regions have important cellular regulatory functions and are associated with regulatory elements, such as promoters, enhancers, silencers, and insulators. This region can produce microRNAs, ribosomal RNAs, and transfer RNAs and form structural elements of the chromosome, such as telomeres, and satellite DNA [72–74]. Now, four strategies have been applied in the AD field to discover genetic variations in the human genome including genetic linkage analyses, candidate gene/pathway association studies, genome-wide association studies (GWAS), and next-generation sequencing (NGS)-based association studies [75].

For example, three mutations in genes, presenilin 2 (PSEN1), presenilin 2 (PSEN 2), and amyloid precursor protein (APP), were suggested to be causative in early-onset familial AD using genetic linkage analyses. Besides, the apolipoprotein E gene (APOE) alleles were identified as risk factors for late-onset AD by candidate gene/

pathway approaches [76]. The International Alzheimer's Disease Project (IGAP) has collected many patients' samples to conduct large GWAS samples of LOADs [77, 78]. This strategy has also confirmed that APOE 4 is the most important genetic risk for AD [64, 75, 79]. To discover the extremely rare variants in AD, NGS was used to complement GWAS which requires large samples. Susceptibility loci, such as *ZNF655*, *ZBTB4*, *TTC3*, *TM2D3*, *PLD3*, *NOS1AP*, *NCSTN*, *IGHG3*, *GRN*, *FSIP2*, *CSF1R*, *CHMP2B*, and *ARSA*, were missed by GWAS but have been found to be related to AD development by NGS in very small population [80–83].

In order to better understand AD etiology, it might be necessary to consider additive or multiplicative effects, as well as the interaction of genes with the environment. However, more information is still needed, such as how genetic variations and environmental risk factors interact and mitochondrial genetic variation.

Diagnosis

Mental capacity and cognition preservation play an important role in the maintenance of autonomy in elderly people. Early detection of pathological cognitive impairment facilitates early, more effective treatment interventions focused on restoration and prevention. However, early detection of cognitive decline is a challenge due to the insidious symptoms, which are usually initially diagnosed as normal age-related cognitive decline [84, 85]. Requiring a longitudinal follow-up using multiple diagnostic criteria, mild cognitive impairment is often misdiagnosed as various different forms of dementia [86].

While magnetic resonance imaging (MRI) and genetic tests are the most common methods used in the clinic for the diagnosis of AD, there is currently no method available to detect early AD. Later-stage diagnostics are usually time- and money-consuming, and most importantly, they are not especially suitable for early detection and screening of large populations in a short time. As a result, an ideal diagnostic tool, sensitive enough to detect early disease, with a high

specificity is needed. AI, being noninvasive and practical, could be an appropriate candidate.

AI can detect prognostic signals from data that can be easily collected such as MRI data, and electronic health records (EHRs). These signals enable the screening of aging populations prospectively. Currently, test results analyzed and interpreted by trained people may lead to delay in diagnosis. However, these delays can be reduced by using the AI approach. There remains much room for improvement, but applying AI to diagnose AD is already growing at an amazing speed [87–89]. For example, one study applied an artificial neural network (ANN) model to the diagnosis of AD within a cohort of 2482 community-dwelling people aged 60+ over 3 years. The study aimed to establish an early warning ANN model with high accuracy and diagnostic efficiency and to find a biomarker which was sensitive to early exploration. This model could be used as a low-cost, practical tool for the early detection and diagnosis of AD [90].

Physicians currently rely on subjective self-reported clinical measurements to diagnose and detect response to therapeutic intervention due to a lack of appropriate biomarkers [91]. Machine learning models can be used to identify the biomarkers of response to treatment from clinical trials. In fact, large publicly funded databases, namely, the ADNI, have set biomarker identification as one of their major objectives. We summarize **possible applications of AI in AD diagnosis and drug development** (Fig. 2).

Therapy

There are no effective treatments for AD; however, the large pharmaceutical companies have slowed their work in this field because of the high failure rate of clinical trials [92–95]. For example, from 2002 to 2012, more than 400 clinical trials for treating AD were performed, but only 1 drug was approved [96]. This highlights the complexity of developing personal drug strategies and provides the opportunity to use new approaches to design and discover new drugs.

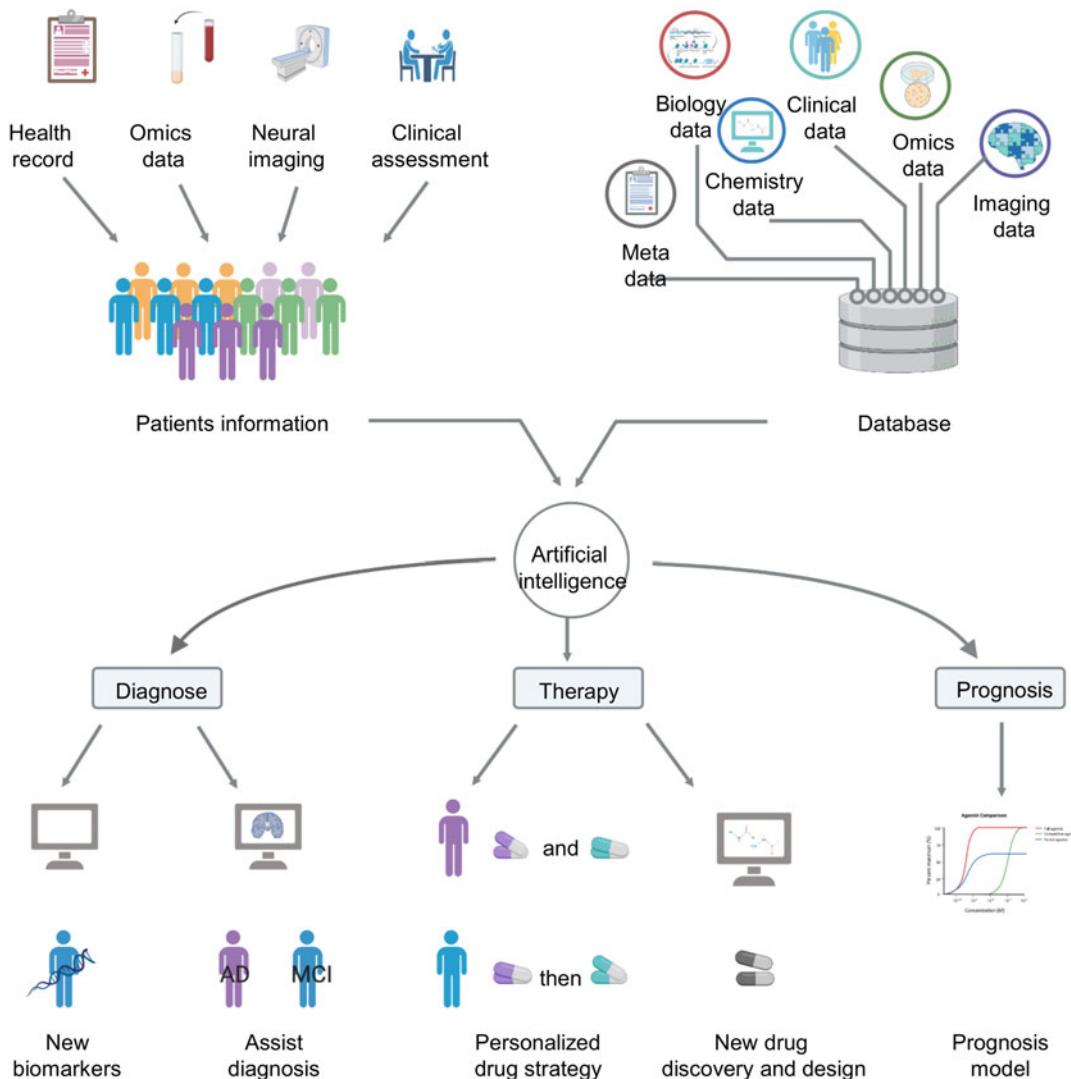


Fig. 2 Possible applications of AI in AD diagnosis and drug development. Through the use of patients' information and big databases, AI could improve diagnosis,

treatment, and prediction of prognosis. For details see section “[Application of Machine Learning in Clinical Work for Alzheimer's Disease](#)”

The use of machine learning techniques for developing personal drug strategies is becoming more and more popular because the whole data of an individual including clinical history, transcriptomic, and neuroimaging as well as biomarker data can be fed into an algorithm of neurodegeneration disease [97]. Using deep learning by unsupervised models may be one approach to stratifying patients to reduce dimensionality in high-dimensional labelled data and to classify patients' outcome. Patients with different

endotypes or subtypes of AD which are not obvious using traditional diagnostic methods may be identified [98]. These subtypes provide evidence for further development of personal drug strategies. For example, using this evidence, physicians may prescribe single drug or combination drugs to an individual patient to improve treatment effects.

AI could also be a new way to design and discover new drugs for AD. As mentioned before, AD pathology involves a vast array of mechanisms. How to explore the data related to these

pathways in an efficient, holistic, and thorough manner is key to understanding AD. However, this can be a challenge for individual researchers. AI can help make sense of or even predict or design new drugs. Knowledge graphs, which link genes, diseases, and drugs, are built from the integration of different data types including ChEMBL, Ensembl, OmniPath, KEGG, and PubMed [99, 100]. This approach can highlight the less-obvious links between drug targets and AD. However, the downside is lack of granularity in biological relationships, which leads to reductions in specific predictions [101]. Several relational inference methods have been published for AD. In contrast to knowledge graphs, machine learning enables a more detailed biological specification. By extracting gene expression data from healthy control and individual patients' gene expression data, molecular networks can visualize biological processes that change in different disease stages. For instance, a combination of Bayesian inference, clustering, and co-regulation was used to analyze transcriptomic data collected from the brain tissue of individuals with late-onset AD and non-AD controls [102]. A group of microglial-specific genes coding the TYROP protein and immune-related genes were found to be highly expressed in late-onset AD patients. After further verifying the function of TYROP in the AD mouse model, researchers found a deficiency of this protein, showing a neuroprotective function [103, 104].

Prognosis

Prognosis is as important as diagnosis because it quantifies the potential disease progression, and therefore how to predict when MCI converts to AD is also a hot research topic. Among the studies, some have focused on the roles of different imaging modalities in the prediction of MCI conversion [105]. Compared to this kind of early study, an atypical approach using multiple modalities was able to concatenate features into a combined feature set that was further used to build up a classifier. Because of the high dimensionality of the combined feature set, researchers built a

logistic regression model with 71.6% accuracy in classifying MCT conversion rate over 4 years [106]. Another line of research which encapsulated each modality feature could better keep intramodal integrity and reveal intramodal differences. One common approach was multiple kernel learning (MKL), which achieved classification with 76.4% accuracy, 81.8% sensitivity, and 66% specificity [107]. Besides using baseline multimodal imaging data alone, some researchers tried to use longitudinal multimodal image data. For example, imaging and Neuropsychological Status Exam Score (NPSEs) data, both at baseline and development stages, were used for AI-based prediction; in this way, the accuracy, sensitivity, and specificity reached 81.40%, 79.69%, and 83.08%, respectively [108]. Finally to increase the accuracy, researchers used florbetapir-PET, together with MRI, FDG-PET, and ADAS-cog scores, which increased accuracy to 86.05% with 81.25% sensitivity and 90.77% specificity [109].

Future Perspectives and Concluding Remarks

As the quality of life and clinical technologies improve in the twenty-first century, a dramatic increase of lifespan, and correspondingly an increased aging population on Earth, is expected. Since the elderly population is more susceptible to disease, infection (e.g., COVID-19), and neurodegenerative diseases (AD), we foresee pressure on society and the healthcare system [110–112]. Thus, it is timely and necessary to apply AI for clinical use. In recent years, AI applications have been widely used in precision medicine, including AI diagnosis, prognosis, and drug development. AI-aided medical applications have not only supported the doctors, researchers, and scientists to increase inefficiency and provide decision-making support but have also accelerated the development of the medical and healthcare industry. With more and more breakthroughs such as DeepMind's AlphaFold 2, we can foresee that AI will be able to revolutionize the drug development and disease diagnosis industry in the near future.

Acknowledgments The authors acknowledge the valuable work of the many investigators whose published articles they were unable to cite owing to space limitations. The authors thank Dr. Chenglong Xie for discussion and Thale Dawn Patrick-Brown for reading the manuscript. E.F.F. was supported by HELSE SØR-ØST (#2017056, #2020001, #2021021), the Research Council of Norway (#262175 and #277813), the National Natural Science Foundation of China (#81971327), Akershus University Hospital (#269901, #261973), the Civitan Norges Forskningsfond for Alzheimers sykdom (for a 3-year Ph.D. fellowship, #281931), the Czech Republic-Norway KAPPA programme (with Martin Vyhálek, #TO01000215), and the Rosa sløyfe/Norwegian Cancer Society & Norwegian Breast Cancer Society (#207819). G.Y. was supported in part by the British Heart Foundation (Project Number: PG/16/78/32402), in part by the Hangzhou Economic and Technological Development Area Strategical Grant (Imperial Institute of Advanced Technology), in part by the European Research Council Innovative Medicines Initiative on Development of Therapeutics and Diagnostics Combatting Coronavirus Infections Award “DRAGON: rapiD and secuRe AI imaging based diaGnosis, stratification, fOllow-up, and preparedness for coronavirus paNdemics” (H2020-JTI-IMI2 101005122), in part by the AI for Health Imaging Award “CHAIMELEON: Accelerating the Lab to Market Transition of AI Tools for Cancer Management” (H2020-SC1-FA-DTS-2019-1 952172), and in part by the UK Research and Innovation (MR/V023799/1). A.R. was also funded by China Scholarship Council [<http://www.csc.edu.cn/>]; the funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Declaration of Interests

E.F.F. has CRADA arrangements with ChromaDex. E.F.F. and G.Y. are consultants to Aladdin Healthcare Technologies. E.F.F. is a consultant to the Vancouver Dementia Prevention Centre and Intellectual Labs. Z.N, X.J and B.T are affiliated with MindRank AI ltd.

References

- Kulikowski CA. Beginnings of artificial intelligence in medicine (AIM): computational artifice assisting scientific inquiry and clinical art – with reflections on present AIM challenges. *Yearb Med Inform*. 2019;28: 249–56.
- Kulikowski CA. An opening chapter of the first generation of artificial intelligence in medicine: the first rutgers AIM workshop, June 1975. *Yearb Med Inform*. 2015;10:227–33.
- Szolovits P, Patil RS, Schwartz WB. Artificial intelligence in medical diagnosis. *Ann Intern Med*. 1988;108:80–7.
- de Dombal FT, Leaper DJ, Staniland JR, McCann AP, Horrocks JC. Computer-aided diagnosis of acute abdominal pain. *Br Med J*. 1972;2:9–13.
- Shortliffe EH, et al. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. *Comput Biomed Res*. 1975;8:303–20.
- Barnett GO, Cimino JJ, Hupp JA, Hoffer EP. DXplain. An evolving diagnostic decision-support system. *JAMA*. 1987;258:67–74.
- Dechter R. Learning while searching in constraint-satisfaction problems. *AAAI-86 Proceedings*. 1986;178–183.
- Yang YJ, Bang CS. Application of artificial intelligence in gastroenterology. *World J Gastroenterol*. 2019;25:1666–83.
- Ljosa V, Sokolnicki KL, Carpenter AE. Annotated high-throughput microscopy image sets for validation. *Nat Methods*. 2012;9:637.
- Sudlow C, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12: e1001779.
- DesRoches CM, et al. Electronic health records in ambulatory care – a national survey of physicians. *N Engl J Med*. 2008;359:50–60.
- Hsiao CJ, et al. Office-based physicians are responding to incentives and assistance by adopting and using electronic health records. *Health Aff (Millwood)*. 2013;32:1470–7.
- Wouters OJ, McKee M, Luyten J. Estimated research and development investment needed to bring a new medicine to market, 2009–2018. *JAMA*. 2020;323: 844–53.
- Zhong F, et al. Artificial intelligence in drug design. *Sci China Life Sci*. 2018;61:1191–204.
- Rester U. From virtuality to reality – virtual screening in lead discovery and lead optimization: a medicinal chemistry perspective. *Curr Opin Drug Discov Devel*. 2008;11:559–68.
- Rollinger JM, Stuppner H, Langer T. Virtual screening for the discovery of bioactive natural products. *Prog Drug Res*. 2008;65(211):213–49.
- Perez-Sianes J, Perez-Sanchez H, Diaz F. Virtual screening meets deep learning. *Curr Comput Aided Drug Des*. 2019;15:6–28.
- Liew CY, Ma XH, Liu X, Yap CW. SVM model for virtual screening of Lck inhibitors. *J Chem Inf Model*. 2009;49:877–85.
- Melville JL, Burke EK, Hirst JD. Machine learning in virtual screening. *Comb Chem High Throughput Screen*. 2009;12:332–43.
- LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436–44.
- Leelananda SP, Lindert S. Computational methods in drug discovery. *Beilstein J Org Chem*. 2016;12: 2694–718.
- Segler MHS, Kogej T, Tyrchan C, Waller MP. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent Sci*. 2018;4:120–31.

23. Klambauer G, et al. Rchemcpp: a web service for structural analoging in ChEMBL, drugbank and the connectivity map. *Bioinformatics*. 2015;31:3392–4.
24. Subramanian A, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*. 2017;171:1437–1452 e1417.
25. Huang SY, Grinter SZ, Zou X. Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions. *Phys Chem Chem Phys*. 2010;12:12899–908.
26. Kinnings SL, et al. A machine learning-based method to improve docking scoring functions and its application to drug repurposing. *J Chem Inf Model*. 2011;51: 408–19.
27. Jimenez J, Skalic M, Martinez-Rosell G, De Fabritiis G. KDEEP: protein-ligand absolute binding affinity prediction via 3D-convolutional neural networks. *J Chem Inf Model*. 2018;58:287–96.
28. Pereira JC, Caffarena ER, Dos Santos CN. Boosting docking-based virtual screening with deep learning. *J Chem Inf Model*. 2016;56:2495–506.
29. Lusci A, Pollastri G, Baldi P. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *J Chem Inf Model*. 2013;53:1563–75.
30. Duvenaud D, et al. Convolutional networks on graphs for learning molecular fingerprints. arXiv preprint arXiv:1509.09292; 2015.
31. Coley CW, Barzilay R, Green WH, Jaakkola TS, Jensen KF. Convolutional embedding of attributed molecular graphs for physical property prediction. *J Chem Inf Model*. 2017;57:1757–72.
32. Tian S, Li Y, Wang J, Zhang J, Hou T. ADME evaluation in drug discovery. 9. Prediction of oral bioavailability in humans based on molecular properties and structural fingerprints. *Mol Pharm*. 2011;8: 841–51.
33. Kearnes S, Goldman B, Pande V. Modeling industrial ADMET data with multitask networks. arXiv preprint arXiv:1606.08793; 2016.
34. Popova M, Isayev O, Tropsha A. Deep reinforcement learning for de novo drug design. *Sci Adv*. 2018;4: eaap7885.
35. Zhavoronkov A, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnol*. 2019;37:1038–40.
36. Angermueller C, Parnamaa T, Parts L, Stegle O. Deep learning for computational biology. *Mol Syst Biol*. 2016;12:878.
37. Yuan Y, Bar-Joseph Z. Deep learning for inferring gene relationships from single-cell expression data. *Proc Natl Acad Sci U S A*. 2019;10:116.
38. Wen B, et al. Deep learning in proteomics. *Proteomics*. 2020;20:e1900335.
39. Kuhlman B, Bradley P. Advances in protein structure prediction and design. *Nat Rev Mol Cell Biol*. 2019;20:681–97.
40. Abriata LA, Tamo GE, Dal Peraro M. A further leap of improvement in tertiary structure prediction in CASP13 prompts new routes for future assessments. *Proteins*. 2019;87:1100–12.
41. Senior AW, et al. Improved protein structure prediction using potentials from deep learning. *Nature*. 2020;577:706–10.
42. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng*. 2018;2:719–31.
43. Lodwick GS. Computer diagnosis of primary bone tumors. A preliminary report. *Radiology*. 1963;80: 273–5.
44. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*. 2017;284:574–82.
45. Setio AA, Jacobs C, Gelderblom J, van Ginneken B. Automatic detection of large pulmonary solid nodules in thoracic CT images. *Med Phys*. 2015;42: 5642–53.
46. Samala RK, Chan HP, Hadjiiski LM, Helvie MA. Analysis of computer-aided detection techniques and signal characteristics for clustered microcalcifications on digital mammography and digital breast tomosynthesis. *Phys Med Biol*. 2016;61: 7092–112.
47. Samala RK, et al. Mass detection in digital breast tomosynthesis: deep convolutional neural network with transfer learning from mammography. *Med Phys*. 2016;43:6654.
48. Esteva A, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115–8.
49. Panwar N, et al. Fundus photography in the 21st century – a review of recent technological advances and their implications for worldwide healthcare. *Telemed J E Health*. 2016;22:198–208.
50. Gulshan V, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316:2402–10.
51. Poplin R, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*. 2018;2:158–64.
52. Abramoff MD, et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Invest Ophthalmol Vis Sci*. 2016;57:5200–6.
53. Ciresan DC, Giusti A, Gambardella LM, Schmidhuber J. Mitosis detection in breast cancer histology images with deep neural networks. *Med Image Comput Comput Assist Interv*. 2013;16: 411–8.
54. Robboy SJ, et al. Pathologist workforce in the United States: I. Development of a predictive model to examine factors influencing supply. *Arch Pathol Lab Med*. 2013;137:1723–32.
55. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Aff (Millwood)*. 2014;33:1123–31.

56. Van Calster B, Wynants L. Machine learning in medicine. *N Engl J Med.* 2019;380:2588.
57. Mamoshina P, et al. Converging blockchain and next-generation artificial intelligence technologies to decentralize and accelerate biomedical research and healthcare. *Oncotarget.* 2018;9:5665–90.
58. Pyrkov TV, et al. Extracting biological age from biomedical data via deep learning: too much of a good thing? *Sci Rep.* 2018;8:5210.
59. Franke K, Ziegler G, Kloppel S, Gaser C, Alzheimer's Disease Neuroimaging Initiative. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters. *NeuroImage.* 2010;50:883–92.
60. Ng A, et al. IL-1beta, IL-6, TNF- alpha and CRP in elderly patients with depression or Alzheimer's disease: systematic review and meta-analysis. *Sci Rep.* 2018;8:12050.
61. Ng TKS, Ho CSH, Tam WWS, Kua EH, Ho RC. Decreased serum brain-derived neurotrophic factor (BDNF) levels in patients with Alzheimer's disease (AD): a systematic review and meta-analysis. *Int J Mol Sci.* 2019;20:257.
62. Ballard C, et al. Alzheimer's disease. *Lancet.* 2011;377:1019–31.
63. Bi C, Bi S, Li B. Processing of mutant beta-amyloid precursor protein and the clinicopathological features of familial Alzheimer's disease. *Aging Dis.* 2019;10:383–403.
64. Freudenberg-Hua Y, Li W, Davies P. The role of genetics in advancing precision medicine for Alzheimer's disease—a narrative review. *Front Med (Lausanne).* 2018;5:108.
65. Gatz M, et al. Role of genes and environments for explaining Alzheimer disease. *Arch Gen Psychiatry.* 2006;63:168–74.
66. Fang EF, et al. Mitophagy inhibits amyloid-beta and tau pathology and reverses cognitive deficits in models of Alzheimer's disease. *Nat Neurosci.* 2019;22:401–12.
67. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med.* 2019;380:1347–58.
68. Webb S. Deep learning for biology. *Nature.* 2018;554:555–7.
69. Zitnik M, et al. Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. *Inf Fusion.* 2019;50:71–91.
70. Lee Y, et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. *J Affect Disord.* 2018;241:519–32.
71. Ku CS, Loy EY, Salim A, Pawitan Y, Chia KS. The discovery of human genetic variations and their use as disease markers: past, present and future. *J Hum Genet.* 2010;55:403–15.
72. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012;489:57–74.
73. Maston GA, Evans SK, Green MR. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet.* 2006;7:29–59.
74. Telenti A, Lippert C, Chang PC, DePristo M. Deep learning of genomic variation and regulatory network data. *Hum Mol Genet.* 2018;27:R63–71.
75. Fenoglio C, Scarpini E, Serpente M, Galimberti D. Role of genetics and epigenetics in the pathogenesis of Alzheimer's disease and frontotemporal dementia. *J Alzheimers Dis.* 2018;62:913–32.
76. Kehoe P, et al. A full genome scan for late onset Alzheimer's disease. *Hum Mol Genet.* 1999;8:237–45.
77. Kunkle BW, et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Abeta, tau, immunity and lipid processing. *Nat Genet.* 2019;51:414–30.
78. Lambert JC, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet.* 2013;45:1452–8.
79. Pimenova AA, Raj T, Goate AM. Untangling genetic risk for Alzheimer's disease. *Biol Psychiatry.* 2018;83:300–10.
80. Beecham GW, et al. Rare genetic variation implicated in non-Hispanic white families with Alzheimer disease. *Neurol Genet.* 2018;4:e286.
81. Bis JC, et al. Whole exome sequencing study identifies novel rare and common Alzheimer's-associated variants involved in immune response and transcriptional regulation. *Mol Psychiatry.* 2020;25:1859–75.
82. Blue EE, et al. Variants regulating ZBTB4 are associated with age-at-onset of Alzheimer's disease. *Genes Brain Behav.* 2018;17:e12429.
83. Cruchaga C, et al. Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. *Nature.* 2014;505:550–4.
84. Deary IJ, et al. Age-associated cognitive decline. *Br Med Bull.* 2009;92:135–52.
85. Petersen RC, et al. Practice parameter: early detection of dementia: mild cognitive impairment (an evidence-based review). Report of the quality standards Subcommittee of the American Academy of neurology. *Neurology.* 2001;56:1133–42.
86. Brodaty H, et al. Operationalizing the diagnostic criteria for mild cognitive impairment: the salience of objective measures in predicting incident dementia. *Am J Geriatr Psychiatry.* 2017;25:485–97.
87. Balota DA, et al. Predicting conversion to dementia of the Alzheimer's type in a healthy control sample: the power of errors in Stroop color naming. *Psychol Aging.* 2010;25:208–18.
88. Patten RV, Fagan AM, Kaufman DAS. Differential cued-Stroop performance in cognitively asymptomatic older adults with biomarker-identified risk for Alzheimer's disease: a pilot study. *Curr Alzheimer Res.* 2018;15:820–7.
89. Silverberg NB, et al. Assessment of cognition in early dementia. *Alzheimers Dement.* 2011;7:e60–76.

90. Wang N, et al. Application of artificial neural network model in diagnosis of Alzheimer's disease. *BMC Neurol.* 2019;19:154.
91. Meyer SM, et al. Optimizing ADAS-cog worksheets: a survey of clinical trial rater s' perceptions. *Curr Alzheimer Res.* 2017;14:1008–16.
92. Cummings J. Lessons learned from Alzheimer disease: clinical trials with negative outcomes. *Clin Transl Sci.* 2018;11:147–52.
93. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. *Nat Biotechnol.* 2014;32:40–51.
94. Stimulus package. *Nat Med.* 2018;24:247.
95. Zwierzyna M, Davies M, Hingorani AD, Hunter J. Clinical trial design and dissemination: comprehensive analysis of clinicaltrials.gov and PubMed data since 2005. *BMJ.* 2018;361:k2130.
96. Cummings JL, Morstorf T, Zhong K. Alzheimer's disease drug-development pipeline: few candidates, frequent failures. *Alzheimers Res Ther.* 2014;6:37.
97. Grollemund V, et al. Machine learning in amyotrophic lateral sclerosis: achievements, pitfalls, and future directions. *Front Neurosci.* 2019;13:135.
98. Maudsley S, Devanarayyan V, Martin B, Geerts H, Brain Health Modeling Initiative. Intelligent and effective informatic deconvolution of "Big Data" and its future impact on the quantitative nature of neurodegenerative disease therapy. *Alzheimers Dement.* 2018;14:961–75.
99. Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys Rev Lett.* 2018;120:145301.
100. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics.* 2018;34:i457–66.
101. Palop JJ, Chin J, Mucke L. A network dysfunction perspective on neurodegenerative diseases. *Nature.* 2006;443:768–73.
102. Zhang B, et al. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell.* 2013;153:707–20.
103. Haure-Mirande JV, et al. Deficiency of TYROBP, an adapter protein for TREM2 and CR3 receptors, is neuroprotective in a mouse model of early Alzheimer's pathology. *Acta Neuropathol.* 2017;134:769–88.
104. Haure-Mirande JV, et al. Integrative approach to sporadic Alzheimer's disease: deficiency of TYROBP in cerebral Abeta amyloidosis mouse normalizes clinical phenotype and complement subnetwork molecular pathology without reducing Abeta burden. *Mol Psychiatry.* 2019;24:431–46.
105. Jack CR Jr, et al. Brain beta-amyloid measures and magnetic resonance imaging atrophy both predict time-to-progression from mild cognitive impairment to Alzheimer's disease. *Brain.* 2010;133: 3336–48.
106. Ritter K, et al. Multimodal prediction of conversion to Alzheimer's disease based on incomplete biomarkers. *Alzheimers Dement (Amst).* 2015;1:206–15.
107. Zhang D, et al. Multimodal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage.* 2011;55:856–67.
108. Hinrichs C, Singh V, Xu G, Johnson SC, Alzheimers Disease Neuroimaging Initiative. Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. *NeuroImage.* 2011;55:574–89.
109. Wang P, et al. Multimodal classification of mild cognitive impairment based on partial least squares. *J Alzheimers Dis.* 2016;54:359–71.
110. Fang EF, et al. A research agenda for ageing in China in the 21st century (2nd edition): focusing on basic and translational research, long-term care, policy and social networks. *Ageing Res Rev.* 2020;64:101174.
111. Mkrtchyan GV, et al. ARDD 2020: from aging mechanisms to interventions. *Aging (Albany NY).* 2020;12:24484–503.
112. Aman Y, et al. The NAD(+)–mitophagy axis in healthy longevity and in artificial intelligence-based clinical applications. *Mech Ageing Dev.* 2020;185: 111194.



Ontological and Connectivity Structure of Disease-Gene Modules in the Human Interactome

Paola Velardi and Lorenzo Madeddu

Contents

Network Medicine: A New Paradigm for the Study of Diseases	1074
An Introduction to Network Medicine	1074
Machine Learning Challenges in Network Medicine	1075
A Network-Based Analysis of Disease Modules Using a Taxonomic Perspective	1077
Construction of the Interactome Taxonomy (I-T)	1078
Taxonomy Alignment and Labeling	1079
Taxonomy Alignment	1079
Comparing Alternative Induced Taxonomies	1080
Interactome Hierarchy (I-T) Labeling	1081
Experimental Set-up	1081
Discussion	1082
Finding Disease Categories with a Corresponding Dense Neighborhood in the Interactome	1082
Finding Unexpected Structural Relations Between Disease Categories	1083
Detection of Nomenclature Errors in Disease-Gene Associations	1084
Conclusions	1085
References	1085

Abstract

The reductionist approach has dominated scientific research for several centuries and has been widely applied in the biomedical field. However, as we move into the realm of complex diseases, this approach fails to provide the insight needed to explain disease pathogenesis. Network medicine is a new paradigm that applies network science, artificial intelligence (in particular, machine learning and graph mining), and systems biology approaches to study a disease as a consequence of physical interactions within a cell. Pieces of evidence in this field show that if a gene or

P. Velardi (✉)
Department of Computer Science, Sapienza University of Rome, Rome, Italy
e-mail: velardi@di.uniroma1.it

L. Madeddu
Dipartimento di Medicina Traslazionale e di Precisione,
Sapienza Università di Roma, Rome, Italy
e-mail: madeddu@di.uniroma1.it; lorenzo.madeddu@uniroma1.it

molecule is involved in a disease, its direct interactors might also be suspected to play some role in the same pathological process. This evidence has lead to formulating the so-called “disease module hypothesis”: genes involved in the same disease show a high propensity to interact with each other. However, the number and structure of disease modules are largely unexplored. The purpose of this study is to systematically analyze the relationship between structural proximity of disease modules and categorical similarity of diseases, by aligning human-curated disease taxonomies with disease taxonomies automatically induced from proximity relations of disease modules within the human-gene interaction network (*interactome*). We propose a large-scale analysis of a vast collection of diseases leveraging a novel network and taxonomy perspective. Our aim is to support clinical studies by obtaining relevant insights to improve our understanding of disease mechanisms at the molecular level.

Network Medicine: A New Paradigm for the Study of Diseases

The purpose of this section is to introduce the main objectives and challenges of a new emerging and interdisciplinary approach to the study of diseases, network medicine (section “[An Introduction to Network Medicine](#)”), and to provide a summary of state-of-the-art network medicine methods based on machine learning and artificial intelligence (section “[Machine Learning Challenges in Network Medicine](#)”). Finally, we present the motivations and objectives of this study (section “[A Network-Based Analysis of Disease Modules Using a Taxonomic Perspective](#)”), whose aim is to gain more insight on categorical relationships between diseases, through the alignment of human-curated disease ontologies with taxonomic structures automatically induced from proximity relations of disease-related genes in the human interactome network.

An Introduction to Network Medicine

In the last decades, academic research and technological developments have supported the evolution of medical knowledge, on the one hand,

providing a continuously growing set of biomedical data and on the other hand, revealing a complexity only perceived until now. In this context, biological networks have become a central hub of multidisciplinary research, to address essential challenges on both diagnostic and therapeutic aspects such as drug development and disease classification [1, 2]. The reductionist approach to scientific discovery has dominated modern Western thought, and has been successful in the quest to understand observable phenomena. This fundamental principle of scientific investigation has also dominated biomedical research, and has led to the identification of causative mechanisms and effective therapeutic approaches for many diseases. Yet, as we move into the realm of increasingly complex, chronic diseases, this straightforward approach fails to provide the insight needed to explain disease pathogenesis. In addition, drug discovery, tied as it is to the purely reductionist principle of single drug target identification and validation, has been waning despite great advances in the fields of structural biology and bioinformatics. Owing to these limitations, alternative approaches have emerged to better understand diseases and therapeutics. Network medicine is one of such approaches. In medicine, the standard reductionist approach tries to identify a single disease by decoupling the complex biological or medical phenomenons into multiple components. Network medicine surpasses the standard reductionist approach, exploiting the network topology and the network dynamics (e.g., the information flow across the network) of biological networks to understand the pathogenic mechanism underlying the complex molecular interconnections. Network medicine is a new research field that applies network science, artificial intelligence, and machine learning to “-omics” (genomic, proteomic, and transcriptomic) data with the aim of studying, preventing, and treating diseases.

The central finding of network medicine is the following: “given the functional interdependencies between molecular components in a human cell, a disease is rarely the consequence of an abnormality in a single gene, but reflects the perturbations (due by mutations, deletions, copy number variations, or expression changes) of the

complex network of intra-cellular interactions” [4]. Network perturbations such as gain and deletion of disease nodes or edges can result in a similar disease phenotype altering the activity of the disease module [3]. Network medicine, therefore, offers a new perspective of the disease with respect to the complex system of molecular interactions within our organism. This new way of approaching the disease has led to a fundamental “holistic” observation and hypothesis of the disease mechanisms, the *disease module hypothesis*. According to Barabási et al. [4]: “A disease module represents a group of network components that together contribute to a cellular function whose disruption results in a particular disease phenotype.” From the network science perspective, molecular components (e.g., proteins) associated with the same disease are not scattered in the human interactome, but tend to interact with each other forming a network local structure or neighborhood, the disease module [2–4]. From a medical and biological perspective, the disease becomes a reflection of the alteration of one or more biological processes that interact in the human interactome [2].

Despite the potential relevance of this hypothesis, due to the high estimated incompleteness of disease-gene associations modeled in the human interactome [5], disease modules are not yet molecularly well defined and devoid of a clear, dense network structure in literature [2].

In addition to the formulation of this “local” network hypothesis of the molecular components associated to a disease, network medicine studies have shown a “global” network tendency for pathobiologically similar diseases to have their respective modules located in adjacent or overlapping areas of the interactome [2, 3, 6]. According to Loscalzo et al. [3] and Menche et al. [2], “proximity and degree of overlap of two disease modules (in the human interactome) has been found to be highly predictive of the pathobiological similarity of the corresponding diseases” and “network-based location of each disease module determines its pathobiological relationship to other diseases.” Indeed, different disease modules can overlap (see Fig. 1), so that perturbations caused by one disease can affect other disease modules that could leading to

comorbidity and pathogenesis mechanisms [4]. A further objective of network medicine is to construct a complete overview of the disease at the molecular level. To accomplish that, recent studies have analyzed the relationships between disease modules and drugs, extending the human interactome with drug data, such as drug-target associations and drug-drug interactions. Similarly to the disease module relationships, recent studies in network medicine proposed and tested a novel network proximity principle regarding the relationships between disease and drug in the human interactome. According to Cheng et al. [1], “a drug with multiple targets to be on-target effective for a disease or to cause off-target adverse effects, its target proteins should be within or in the immediate vicinity of the corresponding disease module in the human interactome.” Furthermore, Cheng et al. [7] expanded the drug-disease network proximity principle to polypharmacy, identifying network topological rules on the adverse or therapeutic impact of drug combinations in treating complex diseases such as hypertension and cancer. Moreover, they propose a drug or drug-target modules perspective, characterized by a well-defined network-based footprint but smaller in size compared to disease modules, moving beyond the “one-drug, one-target” paradigm. In network medicine most of the efforts are focused on using these findings to discover pathobiological relationships underlying molecular networks, such as disease drivers, comorbidity, and pathogenesis mechanisms, to refine the classification of diseases and support drug repurposing and discovery.

Machine Learning Challenges in Network Medicine

Traditional ways to address network medicine tasks are time consuming, labor intensive, and extremely expensive techniques. For example, the assessment of disease-gene associations is performed by statistical studies based on sequencing the DNA of a large number of patients affected by a given disease, known as genome-wide association studies (GWAS) (<https://www.genome.gov/27541954/dna-sequencing-costs-data/>).

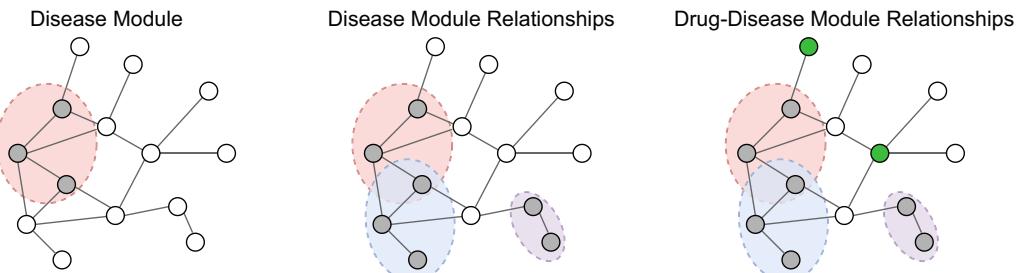


Fig. 1 Disease module properties and observations

Network methods and machine learning have a high potential to boost progress in this field. Recent advancements in machine learning, especially in graph deep learning, led to powerful methods to interpret the complex and interconnected patterns hidden in biological networks, reducing the search space for wet lab experiments, or providing biological insights.

In this section, we summarize some major machine learning challenges in network medicine, and their related issues. As summarized in the previous section, challenges in network medicine ranges from the analysis of the molecular components of a disease to the discovery of therapies based on combinations of drugs. The identification of all disease-associated or causing gene products (i.e., disease genes) is a crucial step to shed light on the nature of the disease modules and their molecular mechanisms. The traditional ways to assess disease-gene associations imply time-consuming and labor-intensive wet lab experiments such as linkage analysis or genome-wide association studies. In recent years, the literature has explored several machine learning strategies to identify disease-gene associations [8]. ProDiGe [9] and CATAPULT [10] use SVM in a positive-unlabeled (PU) learning scheme to rank candidate genes. Other scholars in the field propose deep graph representation methods. In [11], authors predict disease-gene associations by combining node2vec embeddings and features based on network motifs [12]. RW^2 [13] extends node2vec to jointly integrate high-order network structure and disease relationships. Furthermore, the authors of [14] propose to analyze diseases through a deep learning method combining

relation network (RN) and graph convolution neural network (graph CNN) to classify cancer types of patients.

Another relevant field of research is drug repurposing, the identification of new therapeutic effects for a drug [15]. Drug repurposing or repositioning tools are high demanding since the development of a new drug (i.e., drug discovery) is a time-consuming, laborious, expensive, and high-risk process. According to [16], the discovery and development of a new molecular entity (NME) takes 13 years and billions of dollars. Despite these issues, the trend of the number of drugs approved by the Food and Drug Administration (FDA) every year is decreasing [1]. In network medicine, drug repurposing is usually achieved by discovering a new candidate drug for a disease by exploiting network connections between the corresponding drug targets (i.e., proteins) and the disease module. In recent years, literature has explored several machine learning strategies to identify drug-target interactions [17, 18]. In drug-target interaction prediction, DBSI and TBSI [19] are drug or target similarity-based models based on DeepWalk [20]. DTINet [21] infer drug interactions by ranking and learning low-dimensional vector representations of drugs and proteins by combining a network diffusion algorithm and a dimensionality reduction scheme. Other challenging drug-related problems have been recently addressed. For example, [22] assesses new therapeutic combinations of drugs (polypharmacy) using graph convolutional neural networks. In [23], the authors employ the graph convolutional networks to find candidate therapies for the Sars-Cov-2.

Despite the great interest of these approaches, several issues complicate the application of machine learning methods to “-omics” data.

1. Incompleteness: It is estimated that only 20–30% of the existing protein-protein interactions (PPIs) in the human interactome have been discovered [5].
2. Reliability: Protein-protein interaction as well as disease-gene association datasets available in literature are noisy and rely on different evidence scores, based on the reliability of the sources [3, 24].
3. Absence of negative knowledge: When a biological association is not present between two biological entities, we are not sure if it actually does not exist or if it is still unknown [18, 25, 26].

On the other hand, the interest of scientists in network medicine remains high, since a growing interdisciplinary effort gives hope for an acceleration of results in this field.

A Network-Based Analysis of Disease Modules Using a Taxonomic Perspective

Human-curated disease ontologies are widely used for diagnostic evaluation, treatment, and data comparisons over time and clinical decision support. The classification principles underlying these ontologies are historically guided by the analysis of observable, often anatomical, pathological similarities between disorders. Although, thanks to recent advances in molecular biology, disease ontologies are slowly changing to integrate the etiological and genetic origins of diseases, nosology (the science of building medical taxonomies) still reflects a reductionist perspective.

Proximity relationships of disease modules (hereafter DM) in the human interactome network are also increasingly used in diagnostics, to identify pathobiologically similar diseases and support drug repurposing and discovery. Similarity relations induced from structural proximity have also several limitations, although of a different nature w.r.t. those encountered in human-curated ontologies. These

limitations (incompleteness, reliability, and absence of explicit negative knowledge) have been already discussed in previous section “[Machine Learning Challenges in Network Medicine](#).”

The purpose of the study described in this chapter is to shed more light on disease similarities by analyzing the relationship between categorical proximity of diseases in human-curated ontologies and structural proximity of the related DM in the interactome. Our clinical research question is the following: Can we detect meaningful correspondences and discrepancies between distance-based hierarchies automatically generated from disease genes in the interactome network and etiological, human-curated disease taxonomies? The analysis of commonalities and differences among structural and categorical disease relationships could shed more light on the molecular interactions underlying the disease mechanisms, and help refine the human disease classification and better identify limitations and perspectives of current module-based computational approaches to the study of diseases.

Figure 2 shows the workflow of the proposed approach, described in detail in the next sections.

1. **Taxonomy induction:** First, we automatically induce a hierarchical structure of diseases based on proximity relations of DMs in the human interactome. This taxonomic structure is hereafter referred to as the interactome taxonomy (I-T).
2. **Taxonomy alignment and labeling:** Next, we label the intermediate nodes of the I-T using taxonomy alignment and labeling algorithms. The algorithm finds the best map between categorical nodes in a human-curated *reference ontology* (hereafter R-T) and the unlabeled inner nodes of the I-T.
3. **Taxonomy analytics:** The alignment between I-T and R-T supports a large-scale analysis of a vast collection of diseases jointly from an ontological and molecular perspective. We provide insights to refine state-of-the-art nosology and knowledge on disease interactions, by using our framework to investigate the efficacy of the anatomical disease classification principle at the molecular level, identify nomenclature

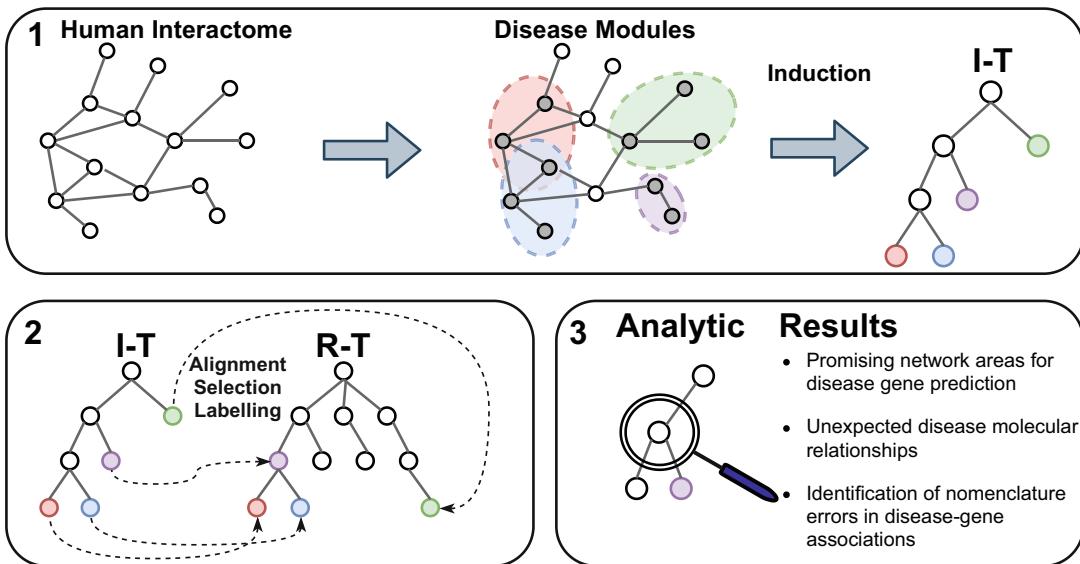


Fig. 2 The work flow of our study. Box 1 shows the taxonomy induction phase, Box 2 represents the phase of taxonomy alignment and labeling, and Box 3 summarizes the results of the analytic phase

error in disease-gene associations, and discover unexpected hidden molecular mechanisms among diseases.

literature to represent disease modules [2, 11]. Given the human interactome network $G = (V, E)$ and a disease d in a set of diseases D_{it} :

- 1. Induced module:** The induced module $I_d = (V_d, E_d)$ is a subgraph of G , where $V_d \subseteq V$ is the set of genes nodes associated with d and E_d is the set of gene-gene interactions $E_d = \{(u, v) | (u, v) \in E \text{ and } u, v \in V_d\}$ [2]. This definition includes in a DM all the disease proteins but, due to the incompleteness of the network, it is usually a graph with many connected components and devoid of a strong local structure [11].
- 2. Largest connected component (LCC) module:** The LCC_d module is the largest connected component of I_d [2, 11]. Unlike the induced module I_d , LCC_d usually has a denser local structure but may not include all the disease-related nodes d .

Given the human interactome G , a set of diseases D_{it} , and their disease modules DM_{it} in G , hierarchical clustering is performed using a distance matrix of disease modules (defined as previously explained), based on the following network-based distance measure (Eq.(1)) [1–3]:

Construction of the Interactome Taxonomy (I-T)

We induce a disease taxonomy (named interactome taxonomy, I-T) by applying hierarchical agglomerative clustering to the human interactome network, exploiting proximity relations of disease modules. Hierarchical agglomerative clustering (HAC) is a set of greedy approaches that create a hierarchy of clusters from unlabeled input data. Given a distance matrix of seed clusters, the HAC algorithm iteratively merges two clusters based on a selected intercluster distance measure. Common methods to merge clusters are average and complete linkage.

In our context, clusters are DM in the human interactome network. However, due to the high incompleteness of the disease-gene associations modeled in the human interactome [27–29], disease modules are not molecularly well defined and devoid of a clear, dense network structure in literature [2]. To cope with this problem, we use two alternative definitions of modules in network science, commonly used in network medicine

$$\text{dist}(A, B) = \frac{\sum_{a \in A} \min_{b \in B} SP(a, b) + \sum_{b \in B} \min_{a \in A} SP(a, b)}{|A| + |B|} \quad (1)$$

where A, B are respectively the set of nodes in modules DM_A and DM_B associated to diseases $d_A, d_B \in D_{it}$, and SP is the shortest path length between two given nodes in G .

In our experiments, we used two DM definitions (induced module and LCC) and two cluster-merge methods (average and complete), and we select the best solution among the resulting four I-Ts, using the methodology described in section “Comparing Alternative Induced Taxonomies.”

Taxonomy Alignment and Labeling

The result of the hierarchical clustering algorithm is a binary tree taxonomy, hereafter referred to as interactome taxonomy (I-T). I-T is a connected directed acyclic graph $T(V_T, E_T)$ in which nodes V_T represent disease concepts and edges represent “is-a” semantic relationships (Edge (v, u) with $u, v \in V_T$ means that v is a kind of u). In our context, leaf nodes (i.e., nodes with out-degree equal to zero) are “specific” diseases D_{it} , physically represented by the corresponding modules DM_{it} , and the inner nodes (i.e., nonleaf nodes) are disease categories DC_{it} (Note that the most generic concept of “disease” is the root node, a node with in-degree equal to zero.). Note that inner nodes $c \in DC_{it}$ are unlabeled, and extensionally defined by the set of their subsumed disease nodes D_{it}^c , also denoted as the C_c clusters.

Similarly, given a “reference” human-crafted taxonomy, denoted as R-T, let $T(V_R, E_R)$ be the set of its nodes and edges, DC_{rt} its disease categories, and $C_{c'}$ the clusters associated with category nodes $c' \in DC_{rt}$.

Taxonomy Alignment

Whatever the choice of the R-T, the R-T and the I-T are expected to be structurally diverse and defined on different sets of diseases nomenclatures, D_{rt} and D_{it} . For example, R-T has usually a polyhierarchical structure, while I-T is a binary tree.

To compare I-T and R-T we first need to create a mapping M from D_{it} to D_{rt} nomenclatures (details on the mapping are given in the experimental section). In Fig. 3, mappings among nodes of the two taxonomies are highlighted using the same colors. As shown, M is not one to one and there are four cases:

1. Case 1: Some D_{it} nodes may map onto inner nodes in the R-T.
2. Case 2: Some D_{it} nodes may map onto multiple nodes in the R-T.
3. Case 3: Some D_{it} nodes may map onto the same node in the R-T.
4. Case 4: Some D_{it} may have no mappings in the R-T and vice versa. These nodes without mappings are the white leaf nodes in both taxonomies.

Our taxonomy alignment procedure consists of three methods: *merge*, *split*, and *pruning*. We apply *merge* and *split* to the R-T to solve cases 1, 2, and 3; instead, *pruning* is applied to the R-T and the I-Ts to solve the case 4. The pseudocode of these algorithms can be found in the online Appendix (http://iim.di.uniroma1.it/pdfs/AIM_in_Genomics_Appendix_Springer_Nature.pdf).

The *merge* algorithm turns in leaves all the R-T colored inner nodes. If the node has noncolored descendants, these are simply removed. Else, the node and its colored descendants are aggregated

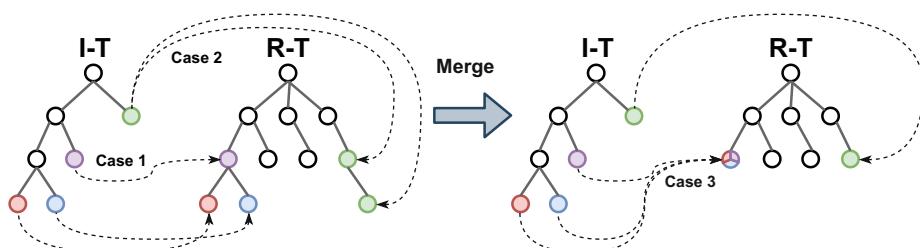


Fig. 3 Visual example of the *merge* algorithm

into a single multicolored node, as shown in Fig. 3 (right).

After the merge, the *split* algorithm splits all the nodes with multiple colors (see Fig. 4). As a result, all colored R-T nodes are all leaf nodes, and every I-T node points to its correspondent R-T node. Note that these new nodes inherit the ancestors of the splitted multicolored node. Furthermore, polyhierarchy in the R-T is preserved, as shown in Fig. 5.

Finally, the *pruning* algorithm (see Fig. 6) prunes both the R-T and the I-Ts by recursively removing survived “white” leaf nodes, those not linked by any mapping relation in M . As a final result, the R-T and the I-T have as leaf nodes the same set of diseases, denoted as D_{\cap} .

Comparing Alternative Induced Taxonomies

As remarked in section “[Construction of the Interactome Taxonomy \(I-T\)](#),” the I-T is built using different definitions of disease modules and different intercluster similarity functions during agglomerative clustering. In this section, we present a method to select the best I-T, among four I-Ts, based on its structural and semantic

closeness with the R-T (Four aligned I-Ts resulting by the combination of the induced and LCC disease module definitions with the average and complete clustering methods.). To this end, we use the Lin semantic similarity [30]:

$$S_T(a, b) := \frac{2 * IC_T(LCS_T(a, b))}{IC_T(a) + IC_T(b)} \quad (2)$$

where IC is:

$$IC_T(x) := -\log \left(\frac{|\text{leaves}_T(x)| + 1}{\text{MaxLeaves}_T + 1} \right) \quad (3)$$

where a, b are two leaf nodes in a taxonomy T , and $LCS_T(a, b)$ is the least common subsumer of a and b in T ; $\text{leaves}_T(x)$ is the set of leaves descendant of x and MaxLeaves_T is the number of leaves in T .

The Lin similarity increases when two nodes are structurally close and decreases otherwise. Furthermore, by construction, the distance between two nodes is normalized with respect of the maximum distance, a property that favors the comparison of taxonomies with different depths. This is a desirable property since the I-T is a binary tree and has a much higher depth than the R-T.

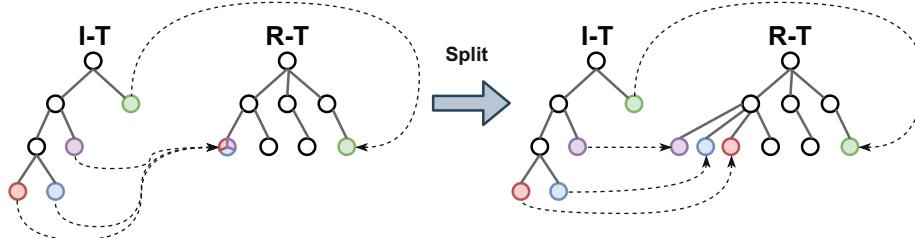


Fig. 4 Visual example of the *split* algorithm

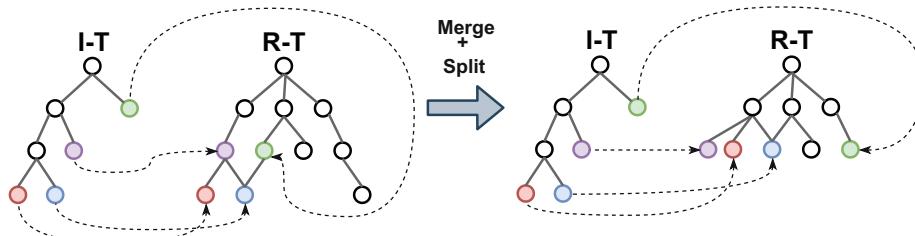


Fig. 5 Visual example of the *merge* and *split* algorithms for a polyhierarchical case

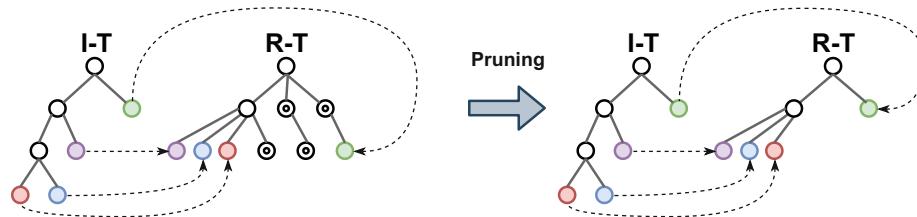


Fig. 6 Visual example of the *pruning* algorithm

To compare I-T and R-T, first, we calculate the pairwise Lin similarity $S_T(d_1, d_2)$ in D_{\cap} for each taxonomy, where $\{(d_1, d_2) | d_1, d_2 \in D_{\cap} \text{ and } d_1 \neq d_2\}$. Next, for each taxonomy, we construct a vector of $S_T(d_1, d_2)$ pairs, and calculate the cosine similarity of R-T and I-T vectors. The idea is that if the two taxonomies are similar, disease pairs that are “close” in one taxonomy should be “close” also in the other taxonomy, and those who are far apart in one taxonomy should be far apart also in the other.

The experimental application of this methodology is described in section “[Discussion](#).”

Interactome Hierarchy (I-T) Labeling

As previously noted, the inner nodes of the aligned I-T have no semantic labels. To facilitate comparison, we defined an algorithm to label each inner node in the I-T with the most similar category label in the R-T. In order to find the most similar R-T category node, we define the *cluster* C_c associated with a category node c in a taxonomy as the set of all its descendant disease nodes that are also in D_{\cap} .

Then, the *labeling* algorithm (see online Appendix for details (http://iim.di.uniroma1.it/pdfs/AIM_in_Genomics_Appendix_Springer_Nature.pdf)) labels every I-T disease category c with the name of the R-T category c' with highest similarity score $sim(C_c, C_{c'})$ between the clusters of c and c' . To compute the similarity between two clusters, we use the Jaccard similarity. Note that the labeling method uses the full set of R-T categories to obtain more fine-grained labels, but the node clusters C_c are defined on the common disease set D_{\cap} of the aligned taxonomies (The

alignment method ends up removing several inner nodes in R-T for the purpose of alignment.).

Experimental Set-up

In this chapter, we considered the human protein-protein interaction network from a recent work of Barabasi et al. [7], which is an extension of a highly cited and popular interactome used by Menche et al. [2] to conduct disease module analysis. The network has $|V| = 16,677$ proteins and $|E| = 243,603$ physical undirected protein interactions.

To construct disease modules, we collected disease-gene associations from DisGeNET [31] with a GDA (GDA is a “reliability” score, for details see www.disgenet.org/dbinfo#section43) score greater or equal of 0.3. Finally, we selected as disease modules the 948 diseases with a set of disease genes of size at least 10.

We selected the disease ontology (DO) as reference taxonomy (R-T) [32]. An alternative widely used reference ontology is ICD-9 (used, for example, in a work by Zhou et al. [33]). However, ICD-9 has been designed to promote international comparability in the classification and presentation of epidemiology and mortality statistics. Its hierarchical structure is not based on etiology, but on anatomical and disciplinary principles. Hence, ICD-9 does not represent a good categorical framework for integrating network-based disease properties. Instead, the disease ontology has been created for describing the classification of human diseases exploiting etiological agents. Its purpose is to identify commonalities of diseases located in a common molecular area (i.e., close to each other in the interactome), originating

Table 1 Comparison of Lin-similarity vectors of the aligned R-T (extracted from the disease ontology) and I-T taxonomies obtained with different methods. The value in

	Induced	LCC
Measure	Complete (RD)	Complete (RD)
Cos. Sim. (%)	43.59 (28.55)	46.33 (29.84)
	39.94 (28.58)	43.7 (29.71)

from a particular cell type or resulting from a common genetic mechanism. Therefore, even though the “localist” view of diseases is still a guiding principle, the DO also exploits the molecular insights of disease phenotypes.

By parsing the DO “obo” file (<http://www.obofoundry.org/ontology/doid.html>), we generated a directed acyclic network hierarchy of 10,012 diseases and disease categories, 10,061 edges and 12 levels. To create a mapping M between the two different nomenclatures, we used partial mappings directly provided in DisGeNET and in the DO, which we further extended with the support of clinicians.

First, we applied the comparison method of section “[Interactome Hierarchy \(I-T\) Labeling](#)” to select the best method to induce the I-T taxonomy, i.e., the one producing a taxonomy with the highest similarity with the selected R-T (namely, the disease ontology). Table 1 shows the result of this comparison. Note that similarity values are compared against those obtained by a random shuffling the disease nodes. Based on the results of Table 1, we select the I-T taxonomy obtained using induced modules to represent DMs, and the average linkage method to merge clusters during hierarchical clustering. This induced taxonomic structure shows a significant similarity with the disease ontology and therefore represents a good basis for our study.

Discussion

Our research hypothesis in this work is that jointly analyzing the structural proximity of disease modules in the human interactome and the semantic proximity of corresponding diseases in human-cured taxonomies could help both refine the classification of human diseases and identify the limitations and perspectives of current module-based computational approaches to the study of

the round brackets represents the average of values generated by ten random distributions of leaf nodes

diseases. In this section, we summarize the major outcomes of a clinical analysis supported by the methodology presented in previous sections. Our analysis is based both on the study of matching and unmatching pairs of R-T and I-T categories.

Finding Disease Categories with a Corresponding Dense Neighborhood in the Interactome

First, we conducted an analysis to reveal in the human interactome large neighborhoods of disease modules associated with disease categories in R-T. Dense neighborhoods of diseases are useful to identify promising disease categories for disease-gene prediction, drug repurposing, and comorbidities detection. To find these large neighborhoods, we verified the existence of topmost disease categories of the R-T with a high overlapping with some internal nodes in the I-T. An R-T disease category c' that is “well-represented” by an I-T category c implies a strong molecular proximity relationship among the diseases in cluster C_c . Symmetrically, this implies that there exists a molecular mechanism that strengthens the classification principle of the R-T category.

We considered the nine disease categories in the first level of the R-T as the most general disease categories. To evaluate the degree of similarity between these R-T categories and their most similar correspondents in the I-T, we used the Jaccard similarity, i.e., the “label score” computed by the labeling algorithm of section “[Interactome Hierarchy \(I-T\) Labeling](#).”

We also calculated the statistical significance of our results by computing the p-value over a random distribution. Table 2 provides an overview of the topmost R-T disease categories and their relationships with I-T categories induced from DM molecular network proximity. In

Table 2 Correspondence among topmost DO categories and the induced taxonomy

R-T (Disease ontology)	Induced I-T
Disease category name (size)	Best label score (P-value)
Disease of cellular proliferation (255)	54.77% ($3.14 \cdot 10^{-20}$)
Disease of anatomical entity (434)	50.05% (0.08)
Genetic disease (12)	41.66% ($6.14 \cdot 10^{-10}$)
Disease by infectious agent (10)	30% ($1.92 \cdot 10^{-7}$)
Physical disorder (21)	26.09% ($1.51 \cdot 10^{-9}$)
Disease of mental health (76)	21.51% ($1.06 \cdot 10^{-13}$)
Syndrome (42)	21.27% ($8.69 \cdot 10^{-11}$)
Disease of metabolism (55)	16.36% ($4.66 \cdot 10^{-11}$)

particular, we found that the R-T disease categories that show a higher localization in a network neighborhood are “disease of cellular proliferation” and “genetic disease.” This means that tumors and genetic diseases are highly localized in two neighborhoods of the human interactome. From a biological network perspective, close DMs of “disease of cellular proliferation” are motivated by the fact that cancer diseases have similar genetic causes in proliferation control genes such as the well-known *P53* [4, 34, 35].

The second best matching category is “disease of anatomical entity,” i.e., disease grouped by human experts according to an anatomical localization principle. However, as shown in the table, the similarity value is high but not statistically significant. This is motivated by the fact that diseases belonging to this topmost category are grouped in diverse subcategories scattered over the network rather than in a large “anatomical” neighborhood. To confirm this hypothesis, we performed a systematic pair-wise comparison among subcategories of “disease of anatomical entity.” We found that very rarely category pairs belonging to different anatomical subsystems have overlapping clusters in the I-T, with some obvious and well-documented relations like nervous and respiratory systems, gastrointestinal and integumentary systems, and musculoskeletal and cardiovascular systems [36–38]. In other terms, the validity of the anatomical classification principle is not disproved by the DM localization hypothesis.

One limitation of the above analysis is that, as remarked in section “Machine Learning Challenges in Network Medicine,” a well-known problem of the human interactome network is its

incompleteness [5]. It follows that, while positive results (disease categories corresponding to highly overlapping disease modules) are useful pieces of evidence to identify interesting areas of the interactome to discover new disease-gene associations, the absence of such evidence could be either motivated by the nonexistence of a similarity relation, or by a lack of knowledge on gene interactions in specific areas of the interactome.

Finding Unexpected Structural Relations Between Disease Categories

A more interesting result would be to identify “unexpected” neighborhoods in the I-T, e.g., disease categories that are not connected in human-curated taxonomies but whose strong molecular similarities suggest that one such connection should be exploited to enrich the ontology.

To help finding these relations we developed a visual tool to explore the labeled I-T in a more systematic way. Clinical experts have identified, among the others, the following interesting results: There exist strong unexpected molecular relationships between glaucoma and pulmonary arterial hypertension (see Fig. 7), cholestasis and chronic obstructive pulmonary disease (COPD), peroxisomal diseases, and ciliopathy-related syndromes. For these relationships, we were able to find confirmations in very recent clinical studies. For example, Gupta et al. [39] and Lewczuk et al. [40] confirm our finding by shedding light on common molecular mechanisms and manifestations between pulmonary hypertension and glaucoma through multiple case reports. Instead,

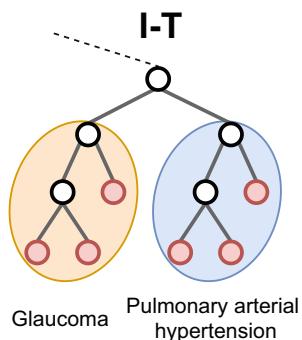


Fig. 7 A branch of the selected I-T groups together glaucoma-related diseases and pulmonary arterial hypertension-related diseases

Tsechkovski et al. [41] observed that cholestasis and COPD pathomechanisms are mediated by common molecular components like the alpha-1 antitrypsin protein. It is interesting to note that there is an emerging hypothesis connecting gut, liver, and lung as playing a key role in the pathogenesis of COPD [42]. Finally, Miyamoto et al. [43] and Zaki et al. [44] found biological mechanisms between peroxisomal diseases and ciliopathy-related syndromes (e.g., Joubert syndrome, Bardet-Biedl syndrome, and Jeune syndrome). These relationships should be used to extend the DO.

Other identified unexpected relationships lack at the moment support from published studies, however the results reported above demonstrate the relevance and potentials of our proposed methodology (Clinical confirmation of our findings is clearly outside the scope of this research, although it represents a study hypothesis for further research by clinicians in the field.).

Detection of Nomenclature Errors in Disease-Gene Associations

Finally, we tried to identify “unconvincing” strong matches between R-T and I-T categories. An in-depth analysis by clinicians has led to the detection of a number of nomenclature errors in DisGeNET.

Disease modules in the interactome have been identified, as discussed in section “[Experimental](#)

[Set-up](#),” using disease-gene associations in DisGeNET, one of the most widely used association databases. Publicly available disease-gene associations databases are manually or computationally curated and some of them integrate other disease-gene collections. However, especially for ambiguous diseases with similar names, all these mechanisms are prone to errors resulting in wrong disease-gene associations. Although in our work we selected only associations with a high GDA score, nomenclature errors might still survive.

The identification of wrong disease-gene associations is of primary importance both for disease-gene discovery and clinical diagnoses. Indeed, on the one hand, disease-gene discovery tools, using wrong disease-gene associations, would make wrong predictions. On the other hand, clinicians usually make and justify diagnoses using the disease-gene associations contained in public databases (as we said, DisGeNET is one among the most widely used resources) leading to wrong diagnoses or therapies for a patient. Here, we demonstrate that our framework may facilitate the detection of wrong disease-gene associations.

To this end, supported by clinicians, we identified a number of R-T disease categories with an unconvincingly high overlapping with I-T inner nodes. Specifically, a clinical expert analyzed all R-T/I-T categories pairs with a Jaccard similarity score greater or equal than 90%. Then, we manually verified the DisGeNET pieces of evidence supporting the related disease-gene associations. We found a number of nomenclature errors, for example: “obstructive lung disease” and “bone remodeling disease” have several wrong disease associations. More in detail, in “obstructive lung disease” the pulmonary emphysema, focal emphysema, panacinar emphysema, and centrilobular emphysema diseases have the same 12 disease-gene associations almost all supported by the same published study, but related only to the pulmonary emphysema or the generic category of emphysema. The same happens in “bone remodeling disease” for the following diseases: osteoporosis, age-related osteoporosis, post-traumatic osteoporosis, and senile osteoporosis.

Conclusions

We believe that the biomedical understanding of diseases is on the edge of a radical change. The disease module hypothesis, with its relevant applications to disease-gene discovery and drug repurposing, is leading the revolution of biomedical research of the future. For these reasons, we deem it fundamental to discover the degree of relationships between the disease module hypothesis and human-curated disease taxonomies. Given the complexity underlying the disease mechanisms, we developed a machine learning methodology to analyze diseases by leveraging, in a novel way, both taxonomic and molecular aspects. We used our methodology to give an useful global view over the human-curated “general” disease categories and their relationships with molecular network proximity. We used our framework to identify large disease network neighborhoods representing promising areas to discover new disease-gene associations, to find “unexpected” disease associations between disease categories that can be used to enrich human-curated ontologies, and finally to detect a number of wrong disease-gene associations in public databases, whose presence could negatively impact on diagnoses and network-based methods. In conclusion, we presented a novel disease module analytic strategy leveraging both the molecular and the taxonomy perspectives, providing new insights into the molecular mechanisms of diseases and a way to refine the human-curated taxonomy.

References

- Cheng F, Desai RJ, Handy DE, Wang R, Schneeweiss S, Barabási A-L, Loscalzo J. Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nat Commun.* 2018;9(1):1–12.
- Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, Barabási A-L. Uncovering disease-disease relationships through the incomplete interactome. *Science.* 2015;347(6224):1257601.
- Loscalzo J, Barabási A-L, Silverman EK. Network medicine: complex systems in human disease and therapeutics, vol. 1. 1st ed. Harvard University Press; 2017.
- Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet.* 2011;12(1):56–68.
- Venkatesan K, Rual J-F, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, Hao T, Zenkner M, Xin X, Goh K-I, et al. An empirical framework for binary interactome mapping. *Nat Methods.* 2009;6(1):83–90.
- Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. The human disease network. *Proc Natl Acad Sci.* 2007;104(21):8685–90.
- Cheng F, Kovács IA, Barabási A-L. Network-based prediction of drug combinations. *Nat Commun.* 2019;10(1):1–11.
- Ata SK, Wu M, Fang Y, Ou-Yang L, Kwoh CK, Li X-L. Recent advances in network-based methods for disease gene prediction. *arXiv preprint arXiv:2007.10848.* 2020.
- Mordelet F, Vert J-P. Prodigie: prioritization of disease genes with multitask machine learning from positive and unlabeled examples. *BMC Bioinform.* 2011;12(1):389.
- Zeng X, Liao Y, Liu Y, Zou Q. Prediction and validation of disease genes using hetesim scores. *IEEE/ACM Trans Comput Biol Bioinform.* 2016;14(3):687–95.
- Agrawal M, Zitnik M, Leskovec J, et al. Large-scale analysis of disease pathways in the human interactome. In: PSB. World Scientific; 2018. p. 111–22.
- Grover A, Leskovec J. node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016. p. 855–64.
- Madeddu L, Stilo G, Velardi P. A feature-learning-based method for the disease-gene prediction problem. *Int J Data Min Bioinform.* 2020;24(1):16–37.
- Rhee S, Seo S, Kim S. Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype classification. In: Proceedings of the twenty-seventh international joint conference on artificial intelligence, IJCAI-18. International Joint Conferences on Artificial Intelligence Organization; 2018. p. 3527–34.
- Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, Doig A, Guilliams T, Latimer J, McNamee C, et al. Drug repurposing: progress, challenges and recommendations. *Nat Rev Drug Discov.* 2019;18(1):41–58.
- Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL. How to improve R&D productivity: the pharmaceutical industry’s grand challenge. *Nat Rev Drug Discov.* 2010;9(3):203–14.
- Ezzat A, Wu M, Li X-L, Kwoh C-K. Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey. *Brief Bioinform.* 2019;20(4):1337–57.
- Lotfi Shahreza M, Ghadiri N, Mousavi SR, Varshosaz J, Green JR. A review of network-based approaches to drug repositioning. *Brief Bioinform.* 2018;19(5):878–92.

19. Zong N, Kim H, Ngo V, Harismendy O. Deep mining heterogeneous networks of biomedical linked data to predict novel drug–target associations. *Bioinformatics*. 2017;33(15):2337–44.
20. Perozzi B, Al-Rfou R, Skiena S. Deepwalk: online learning of social representations. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. 2014. p. 701–10.
21. Luo Y, Zhao X, Zhou J, Yang J, Zhang Y, Kuang W, Peng J, Chen L, Zeng J. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat Commun*. 2017;8(1):1–13.
22. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*. 2018;34(13):i457–66.
23. Gysi DM, Valle ID, Zitnik M, Ameli A, Gan X, Varol O, Sanchez H, Baron RM, Ghiaian D, Loscalzo J, et al. Network medicine framework for identifying drug repurposing opportunities for covid-19. arXiv preprint arXiv:2004.07229. 2020.
24. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, García-García J, Sanz F, Furlong LI. Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res*. 2016;45:D833.
25. Blohm P, Frishman G, Smialowski P, Goebels F, Wachinger B, Ruepp A, Frishman D. Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic Acids Res*. 2014;42(D1):D396–400.
26. Sachdev K, Gupta MK. A comprehensive review of feature based methods for drug target interaction prediction. *J Biomed Inform*. 2019;93:103159.
27. Venkatesan K, Rual J-F, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, Hao T, Zenkner M, Xin X, Goh K-I, Yildirim MA, Simonis N, Heinzmann K, Gebreab F, Sahalie JM, Cevik S, Simon C, de Smet A-S, Dann E, Vidal M. An empirical framework for binary interactome mapping. *Nat Methods*. 2009;6:83–90.
28. Stumpf MP, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, Wiuf C. Estimating the size of the human interactome. *Proc Natl Acad Sci*. 2008;105(19):6959–64.
29. Luck K, Kim D-K, Lambourne L, Spirohn K, Begg BE, Bian W, Brignall R, Cafarelli T, Campos-Laborie FJ, Charlotteaux B, et al. A reference map of the human binary protein interactome. *Nature*. 2020;580(7803):402–8.
30. Lin D, et al. An information-theoretic definition of similarity. *ICML*. 1998;98:296–304.
31. Piñero J, Ramirez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, Furlong LI. The disgenet knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res*. 2020;48(D1):D845–55.
32. Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, Mungall CJ, Binder JX, Malone J, Vasant D, Parkinson H, Schriml LM. Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res*. 2015;43(D1):D1071–8.
33. Zhou X, Lei L, Liu J, Halu A, Zhang Y, Li B, Guo Z, Liu G, Sun C, Loscalzo J, et al. A systems approach to refine disease taxonomy by integrating phenotypic and molecular networks. *EBioMedicine*. 2018;31:79–91.
34. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. The human disease network. *Proc Natl Acad Sci*. 2007;104(21):8685–90.
35. Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, et al. The genomic landscapes of human breast and colorectal cancers. *Science*. 2007;318(5853):1108–13.
36. Chhabra S, De S. Cardiovascular autonomic neuropathy in chronic obstructive pulmonary disease. *Respir Med*. 2005;99(1):126–33.
37. Huang BL, Chandra S, Shih DQ. Skin manifestations of inflammatory bowel disease. *Front Physiol*. 2012;3:13.
38. Maron BJ, Maron MS. Hypertrophic cardiomyopathy. *Lancet*. 2013;381(9862):242–55.
39. Gupta I, Haddock L, Greenfield DS. Secondary open-angle glaucoma and serous macular detachment associated with pulmonary hypertension. *Am J Ophthalmol Case Rep*. 2020;20:100878.
40. Lewczuk N, Zdebik A, Boguslawska J, Turno-Krecicka A, Misiuk-Hojlo M. Ocular manifestations of pulmonary hypertension. *Surv Ophthalmol*. 2019;64(5):694–9.
41. Tsechkovski M, Boulyjenkov V, Heuck C. A1-antitrypsin deficiency: memorandum from a who meeting> I. *Bull World Health Organ*. 1997;75(5):397–415.
42. Young RP, Hopkins RJ, Marsland B. The gut–liver–lung axis. Modulation of the innate immune response and its possible role in chronic obstructive pulmonary disease. *Am J Respir Cell Mol Biol*. 2016;54(2):161–9.
43. Miyamoto T, Hosoba K, Itabashi T, Iwane AH, Akutsu SN, Ochiai H, Saito Y, Yamamoto T, Matsuura S. Insufficiency of ciliary cholesterol in hereditary zellweger syndrome. *EMBO J*. 2020;39:e103499.
44. Zaki MS, Heller R, Thoenes M, Nürnberg G, Stern-Schneider G, Nürnberg P, Karnati S, Swan D, Fateen E, Nagel-Wolfrum K, et al. Pex6 is expressed in photoreceptor cilia and mutated in deafblindness with enamel dysplasia and microcephaly. *Hum Mutat*. 2016;37(2):170–4.



AIM in Genomic Basis of Medicine: Applications

78

Mayumi Kamada and Yasushi Okuno

Contents

Introduction	1087
Classification of Genomic Variation	1088
Variants in the Coding Region	1088
Variants in the Non-coding Region	1089
Interpretation of Variants Using NLP	1090
Diagnosis (Phenotyping)	1091
Proposal for Optimal Drug Treatment	1092
Summary and Future Implications	1093
References	1094

Abstract

Genomic medicine, which can be used for the appropriate diagnosis, treatment, and prevention of a disease based on information obtained from genome analysis, is now being used in clinical practice to realize precision medicine. It involves the interpretation of genomic variants to clarify their effect on molecular processes, link them to phenotypes, and search for clinical and biological evidence for their relationship with disease. These tasks require an enormous amount of time and effort, which is

why artificial intelligence (AI) is expected to be used. Moreover, it is important to determine which drugs are effective for which genotypes to make treatment decisions. However, it is not realistic to experimentally verify the effectiveness of drugs for various genotypes. Therefore, AI is also being applied to the search for effective drugs for each genotype, which is an important issue in genomic medicine. In this chapter, we introduce the applications of AI to address these important issues in genomic medicine.

Introduction

The Human Genome Project, which started in 1990, revealed that the human genome consists of approximately 3 billion bases, and the reference

M. Kamada · Y. Okuno (✉)

Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto University, Kyoto, Japan
e-mail: kamada.mayumi.2c@kyoto-u.ac.jp;
okuno.yasushi.4c@kyoto-u.ac.jp

sequence of the human genome was determined in 2003. The determination of the reference sequence made it possible to compare the genomes of individuals and discuss differences. In general, approximately 0.1% of the total genome, or 3 million bases, differs between individuals. It is becoming clear that these differences are involved in disease onset, chronicity, and drug efficacy.

Especially in cancer, the response rate to drug treatment differs depending on the genetic differences between cancer cells. For example, although the response rate to standard treatment with drugs is 27.5% [1] in Japanese patients with advanced non-small-cell lung cancer, the response rate increases dramatically to 76.4% [2] in patients with epidermal growth factor receptor variants. Additionally, many molecular-targeted drugs that target overexpressed proteins or those with driver mutations have been developed. It is expected that the optimal drug for patients with cancer can be selected using the presence or absence of somatic variants as an indicator. It has also been reported that the risk of developing various diseases, such as lifestyle-related diseases, Alzheimer's disease, intractable diseases, and cancer, differs depending on the genotype. Thus, "genomic medicine," which can be used for the optimal diagnosis and treatment of a disease based on the patient's genomic information, has begun to be implemented in clinical settings to realize precision medicine.

In genomic medicine, genomic variations detected by DNA sequencers are clinically interpreted based on the patient's background and the vast amount of existing knowledge; then, diagnosis and treatment decisions are made. This interpretation process is referred to as "curation," which requires the aggregation and comprehensive understanding of an enormous amount of information. Although the cost of genome analysis itself is now less than 1000 dollars, that of curation is nearly 100 times more expensive, which is one of the reasons why the use of artificial intelligence (AI) is required. In this chapter, we will introduce the application of AI to genomic medicine, focusing on the following four topics:

1. Classification of genomic variation
2. Interpretation of variants using natural language processing (NLP)
3. Diagnosis (phenotyping)
4. Proposal for optimal drug treatment

Classification of Genomic Variation

A genomic variant is different from a reference sequence at a chromosomal position. Experimental functional validation is required to determine how a variant affects molecular processes. However, with the vast number of variants detected, it is difficult to identify the effects of all of them. Therefore, we can make predictions for de novo genomic variants using AI and the life science/medical knowledge available.

Variants in the Coding Region

Proteins are produced, through transcription and translation, based on the genetic information stored in DNA. After transcription, the region of mRNA that is the target of translation is called "the coding region." For the process of translation, each amino acid that constitutes a protein is designated by three bases (a codon); multiple codons can produce the same amino acid. The amino acid designated by the codon may change because of a change in one of the bases, called a non-synonymous substitution (missense variant). Many computational methods have been developed to predict the relationship between the molecular function of a missense variant and disease.

Various features have been utilized to predict the impact of variants on molecular function. In particular, sequence conservation in homologous sequences and the physicochemical property of protein structure are often used. These features can be used as descriptors in machine learning for more accurate prediction, such as PolyPhen2 [3] using naive Bayes classifiers, MutationTaster [4], and MetaSVM [5] using support vector machine (SVM) [6].

Furthermore, the ensemble (meta-prediction) method, which integrates the scores of each method to make predictions, has shown good accuracy. The first proposed ensemble method, CONDEL [7], uses linear classification to make predictions based on the weighted average scores from five different methods. Following this, many ensemble methods have been developed. REVEL [8] and ClinPred [9] are two examples of successful prediction tools with high accuracy that use random forests. REVEL is also recommended in guidelines for hearing loss [10]. Combined annotation-dependent depletion (CADD) has been claimed to be useful for the diagnostic classification of rare diseases [11]. CADD is an ensemble method that uses logistic regression (version 1.0, SVM-based) and predicts the impacts of variants based on approximately 60 prediction results, including those based on conservation, epigenetic modification, functional prediction, and genetic context [12]. DANN is an extension of this model using deep learning [13].

It is difficult to determine whether molecular functions are genuinely unaffected by a missense variant. Therefore, PrimateAI [14], published in 2018, sequenced six nonhuman primate species and constructed a dataset that shows that most commonly retained missense variants are clinically benign in humans. A deep neural network model trained on this dataset predicted variants associated with rare diseases with high accuracy and detected novel variants associated with intellectual disability. This primate dataset holds great promise for the interpretation of variants of uncertain significance.

Many of the prediction scores of these tools have been published in the dbNFSP database [15].

Epistasis refers to the interactions between different gene loci that affect a single phenotype. Several prediction models based on epistasis have been developed. For example, EVmutation [16] quantifies the effect of multiple mutations on multiple sequence alignments (MSAs) by evaluating the interdependence of amino acid residues between positions using a stochastic model. The validation tests showed that this method can predict the effect more accurately than the method that evaluates variants individually.

EVmutation can evaluate the cooperativity between mutant pairs. It has been extended to a deep generative model called DeepSequence to evaluate higher-order dependencies [17]. In DeepSequence, a variational autoencoder (VAE) is constructed as a generative model with evolutionary statistics obtained from MSA as latent variables. It can predict the effect of variants better than the pairwise model. Various prediction methods using protein sequence generation models have been proposed, including methods using VAE [18–20] and generative adversarial networks [21, 22].

Variants in the Non-coding Region

Non-synonymous variants predict functional consequences at the protein level. Regions that are translated into proteins account for less than 2% of the human genome, and most of the human genome (98%) is composed of non-coding regions that are not translated into proteins. The non-coding regions play a significant role in gene regulation, and some variants in these regions have been reported to be associated with disease [23–26].

The non-coding regions include intronic and exonic splicing enhancers, silencers, insulators, and DNA interactions that influence gene splicing. Thus, it is not easy to identify the effect of a variant on the non-coding region because of its complexity.

Among non-coding variants, many variants in the regions involved in splicing have been detected. Splicing is the process of combining the exon regions of the mRNA precursors transcribed from DNA, excluding the regions that are not involved in determining the amino acid sequence of the protein, called “introns.” One-third of the alleles responsible for hereditary diseases are believed to alter the splicing process [27].

Various methods have been proposed to predict the effects of splice variants. SpliceAI [28] employs a deep neural network architecture consisting of 32 dilated convolutional layers. It utilizes a long window size (10 kbp) and can

predict the effect of a splice variant directly from exon-intron junction sequence data.

Deep learning-based sequence analyzer (DeepSEA) [29] is a multitask hierarchically structured convolutional neural network (CNN) trained on large-scale functional genomic data. It directly learns a regulatory sequence code from large-scale chromatin-profiling data, including data on transcription factor binding, DNase I sensitivity, and histone-mark profiles, and is able to predict the effects of a non-coding variant on these regulatory elements. Utilizing the DeepSEA framework, some non-coding variants have been identified as candidates related to autism spectrum disorder onset [30]. The ExPecto algorithm [31], an extension of DeepSEA, can predict the tissue-specific transcriptional effects of variants from a genomic sequence.

Interpretation of Variants Using NLP

The American College of Medical Genetics (ACMG) guideline for the clinical interpretation of variants has been referred to worldwide for monogenic and other genetic diseases [32]. This guideline was jointly published in 2015 by the ACMG, Association of Molecular Pathology, and College of American Pathologists.

It evaluates multiple pieces of information related to a variant's pathogenicity according to the strength of the evidence and synthesizes the obtained evaluations to determine the final pathogenicity. This guideline allows for a more objective and accurate evaluation.

Some of the information used in the guideline, such as allele frequencies in a population or predictive scores from the tools described above, can be easily obtained from databases and other sources. However, other information, such as functional analysis, case reports, and household information, which is strong evidence of pathogenicity, is often only basically described in the literature and electronic health records (EHRs). Therefore, it is necessary to survey the literature to obtain these evidences.

This evidence search requires considerable time and effort. To search medical papers, a

string-searching algorithm, using relevant keywords as the input, can be used. Currently, the relevant papers must be identified among a vast number of candidate papers that match the keywords. Then, the portion of text that contains the desired information must be located. As more than 200,000 cancer-related papers are published annually, this is a difficult and time-consuming task.

There are high expectations for AI to conduct an efficient evidence search. In an AI-based evidence search, NLP techniques are used to extract genes, drugs, and diseases using entity recognition, targeting the abstracts or full texts of articles. The next step is to extract the relationships between the tagged entities. At this time, a vast number of specific dictionaries and corpora are required to handle technical terms. For this reason, curation is often conducted by corporations and large groups.

The Clinical Interpretations of Variants in Cancer (CIViC) database provides information on all types of variants in cancer and their biological and clinical interpretations (<https://civicdb.org/>). CIViC is unique because it is an expert-crowdsourced knowledge base with researcher (user) participation [33].

More recently, the CIViCmine project has been implemented to improve the efficiency of the curation process. In this project, an information extraction method was developed to obtain evidence on cancer biomarkers that had not yet been aggregated in the knowledge base [34]. Cancer type, gene, drug, and specific evidence type (diagnostic, predictive, predisposing, or prognostic) were extracted from the PubMed Central Open Access subset. As a result, 87,412 gene-cancer associations with clinical relevance were identified. The curators evaluated the obtained biomarkers based on accuracy, usefulness, and necessity, and 73% of the biomarkers were evaluated as useful.

Bidirectional Encoder Representations from Transformers (BERT) [35] is a pre-training language representation method that excels in an array of NLP tasks. Domain-specific BERT models have been released, such as SciBERT [36] for scientific text, BioBERT [37] for

biomedical text, and PubMedBERT for biomedical NLP [38]. As the validation of PubMedBERT has shown that domain-specific models are useful for biomedical NLP tasks, it will be possible to use them for highly accurate information extraction in the future.

Diagnosis (Phenotyping)

The purpose of genomic medicine in clinical practice is to diagnose and predict disease risk. For a long time, it has been believed that facial features and expressions show signs of various genetic diseases, and they are expected to be utilized as diagnostic tools in the medical field. However, because of the wide variety of genetic disorders, it has not been easy to accurately link a disorder with a human facial feature or expression.

DeepGestalt [39] is a CNN-based model that can detect genetic diseases from face pictures with high accuracy. DeepGestalt is trained with approximately 17,000 face pictures of patients with more than 200 genetic diseases and predicts candidate genetic diseases that they may have. The evaluation results showed that accuracy for correct disease was present among 10 predicted candidates in 90.6% of cases (top 10 score), among 5 candidates in 85.4% of cases (top 5 score), and among 1 candidate in 61.3% of cases (top 1 score). In several experiments comparing these answers with those of medical specialists, the diagnostic accuracy of DeepGestalt exceeded that of the human.

PEDIA [40] is an SVM-based phenotype prediction model that combines CADD scores for exome variants and scores for symptoms as features in addition to the DeepGestalt scores described above. This strategy achieved higher accuracy than DeepGestalt alone (86–89% top 1 score) in 679 individuals with 105 different monogenic disorders.

To predict genetic diseases based on phenotypes obtained from EHR data and other sources, Bastarache et al. [41] defined phenotype risk scores, named PheRS, by mapping the clinical features of monogenic diseases to phenotypic information obtained from EHRs. Although this

score is simple, it can discriminate monogenic diseases using phenotypic information with high accuracy. Furthermore, by applying this method to large-scale genomic data, they successfully identified rare variants associated with hereditary diseases that have not been previously recognized.

These reports indicate that genetic features in phenotypic information can be identified using AI. Thus, there is a growing interest in using AI to diagnose genetic diseases based on phenotypic information.

Currently, the use of AI to extract useful information for diagnosis from the results of genome analysis is being investigated on a large scale in many countries. Additionally, for hereditary diseases, such as Mendelian genetic diseases, research on the use of AI to predict pathogenicity and disease risk [42] is being conducted.

For common complex diseases, the polygenic risk scores (PRS) have been widely studied for risk prediction (stratification). The PRS reflects the genetic risk of developing a disease by combining the genotype information of many single nucleotide polymorphisms (SNPs) and has the potential to screen (stratify) people at high risk for a specific disease. The PRS has generally been derived using only summary statistics from genome-wide association studies (GWAS).

Recently, it was reported that breast cancer risk is better predicted from the large-scale data of the UK Biobank using the penalized regression model than using conventional methods [43]. A study predicted the risk of diseases, such as heart attack and diabetes, from the same biobank data using Lasso [44]. This study reported that a higher area under the curve can be achieved by considering more data, such as sex and age, and adding SNP data. In this context, a computational framework for PRS calculation has been developed [45].

However, all of these studies were based on linear regression models and limited to association analyses. To elucidate the complex pathogenesis of multifactorial diseases, it is necessary to consider the combination of multiple variants and the patient's background, such as lifestyle. In the future, the discovery of potential mechanisms from complex interactions is expected using AI

to consider a large number of genetic factors, environmental conditions, and disease phenotype data that are being accumulated.

Proposal for Optimal Drug Treatment

Genomic medicine aims to stratify patients for optimal treatment based on their genetic background. Various deep learning approaches have been developed for optimal drug treatment, especially for cancer.

Several large-scale anticancer drug screenings have been conducted, and the results have been published in public repositories. In particular, projects such as Genomics of Drug Sensitivity in Cancer (GDSC) [46] and Cancer Cell Line Encyclopedia (CCLE) [47] are often used to predict drug sensitivity in cancer. CCLE and GDSC provide genomic data, including somatic variant, copy number variation, and mRNA expression data, to characterize cancer cell lines.

DeepDSC [48] uses a stacked deep AE for feature extraction from gene expression data. Then, precomputed Morgan fingerprints for the compounds are combined with the gene expression features. These are finally used as inputs for the fully connected network model, and the model predicts drug sensitivity (half-maximal inhibitory concentration [IC_{50}]).

MOLI [49], a multi-omics late integration method based on deep neural networks, uses gene expression, copy number, and somatic mutation data as inputs. For each data type, a distinct encoding subnetwork is constructed, and the features are extracted. The obtained features are integrated and input into the final subnetwork to predict drug sensitivity. Validation has shown that MOLI is useful for the use of multiple omics data, and a more accurate prediction can be achieved by integrating different data types after representation learning.

DeepDR [50], by Chiu et al., enables transfer learning by reusing parts of a network trained on other datasets. They prepared encoders that were trained on the Cancer Genome Atlas tumor data ($n = 9059$) in the Genomic Data Commons Portal [51] and performed feature encoding for gene

expression and somatic mutation data from GDSC and CCLE. The two subnetworks were combined with a five-layer feed-forward neural network to predict the IC_{50} . Although it does not utilize compound information such as fingerprints, using pre-trained AEs, it achieves more accurate predictions than linear regression models or SVM.

The above models were combined into a single network for prediction. However, it has been reported that the combination of several independent machine learning models can slightly improve the accuracy of drug response prediction [52].

CDRScan [53] separately constructs five CNN models (each called a base model) and uses them as an ensemble to achieve highly accurate and robust predictions. Four of the five base models have a two-step convolutional architecture, in which each mutation and compound datum is separately convoluted. These two convoluted features are merged (called “virtual docking”), and convolution is applied again. The final model predicts the IC_{50} values. The validation results confirm the usefulness of the proposed architecture, and that the two-step convolutional model is superior to the ensemble approach. However, in terms of robustness and generality, the ensemble approach may be more useful.

In the treatment of cancer, acquired resistance mutations often result in decreased efficacy. For this reason, it is common to administer two or more drugs in combination. Several computational methods have been developed to predict drug synergy.

DeepSynergy [54] is the first to use deep learning to predict the effects of drug combinations. DeepSynergy uses the chemical descriptors for drugs and genomic information of the cell line (gene expression profile for the untreated cell) as inputs to predict drug synergy. To train the model, the developers utilized large-scale oncology screening data produced by Merck & Co. [55] and omics data from GDSC. The Merck & Co. dataset comprises 23,062 samples; each sample consists of 2 compounds and a cell line. It covers 583 distinct combinations, and each was tested against 39 human cancer cell lines derived

from 7 different tissue types. DeepSynergy performed well on validation data. However, the authors pointed out that, as with other methods, it is difficult to predict the interaction of cell lines and compounds that are not included in the training dataset.

Xia et al. [56] utilized more diverse omics data and constructed individual encoding subnetworks, such as the drug sensitivity model described above, to predict combination effects. The model was trained on the large National Cancer Institute-A Large Matrix of Anti-Neoplastic Agent Combinations (NCI-ALMANAC) [57]. The NCI-ALMANAC consists of the therapeutic activity of over 5000 pairs of Food and Drug Administration-approved cancer drugs against a panel of 60 well-characterized human tumor cell lines (NCI-60). The model's input data types are drug descriptors, gene expression, microRNA, and proteomics data. Encoding subnetworks for each input data are learned, and the final prediction subnetwork estimates the growth inhibition value for each drug combination and cell line. Their model showed high performance in the fivefold cross-validation of the dataset. However, as discussed with DeepSynergy, the prediction performance of unknown datasets has not been shown.

Although they did not use genomic profiles, Jiang et al. [58] utilized graph convolutional networks (GCNs) to predict the effects of combinations on each cell line. They constructed a multimodal graph for each cell line using a drug-drug synergy network, drug-target interaction network, and protein-protein interaction network. The constructed multimodal graph was used as the input for the GCN, and the GCN model predicted cell line-specific drug combination synergy.

Here, we have introduced AI approaches to cancer genomics as an optimal drug selection method based on genomic profiles. Moreover, in terms of genomic drug discovery (the development of optimal drugs based on the genome), various major pharmaceutical companies have commenced collaborative research and development on AI. For example, GlaxoSmithKline has invested in the consumer genetics company

23andMe, utilizing the datasets they hold to develop drug targets with AI.

Summary and Future Implications

This chapter has introduced the applications of AI in genomic medicine by addressing the classification of variants, efficiency of curation, and drug selection based on genomic background. The mechanisms of genomic aberrations in many diseases are still unclear, and research is advancing at an ever-increasing pace. In this section, we will discuss the challenges of using AI in genomic medicine.

The first is the amount of data and bias involved. It is necessary to consider genetic divergence in the datasets used for AI training. Most of the genomic information accumulated so far is biased toward the Western population. For example, DeepGestalt has shown low accuracy in the identification of Down syndrome in individuals of African descent compared with those with European ancestry [59]. It has also been reported that PRS prediction in GWAS analysis is prone to uneven performance owing to ethnic group bias in the training data [60].

Additionally, there is limited clinical information on most of the shared variant data. For accurate clinical interpretation (discrimination), it is necessary to use detailed clinical and patient background data as training data. Currently, large-scale biobank cohorts, such as the UK Biobank [61] and the All of Us Research Program [62], are being promoted globally. By utilizing these data, learning data can be made more enrich and reliable. AI is expected to be more accurate and useful in clinical practice in the future.

The second issue is related to ethical issues. As mentioned above, sharing detailed data is essential for the development of a more beneficial AI. However, the amount of detailed genomic and clinical information that can be shared is limited because of restrictions on personal information sharing. The ethical, legal, and social implications of data sharing have been discussed in many countries. The Global Alliance for Genomics and Health, an international community

established to promote the sharing of genomic and clinical data internationally, discusses technical and ethical issues in data sharing. In terms of data utilization in AI algorithm development, the authors proposed data visiting [63], in which researchers apply their algorithms to large-scale studies and clinical cohorts instead of downloading shared data. To this end, discussions on standardization for interoperable and federated genomic and clinical data are underway.

The final topic we will discuss is AI interpretation. The use of AI on vast amounts of data is expected to reveal the latent mechanisms of disease. For clinical application, the interpretability of results is essential because of the high risk of applying predictive results. AI is often described as a black box because the process of making predictions is not visible. It is necessary to clarify the learning process to interpret the prediction results. In deep learning, techniques for visualizing the learning process have been developed. Practical approaches have been proposed to overcome image discrimination problems [64]. However, complex interdependencies exist between the genome and disease; thus, accurate AI output interpretation becomes very difficult [65].

However, currently, techniques are being developed to extract the factors involved in the differences between populations by utilizing Bayesian networks that estimate causal relationships. This technique makes it possible to visualize the crucial factors and is expected to be useful in interpreting interactions [66].

Many methods described in this chapter are still in the research stage, and only a few methods have been put to practical use. Additionally, AI predictions can only be used as a reference for clinical applications. Some AI settings, such as thresholds in prediction, require the judgment of researchers with domain knowledge. Moreover, it is also essential to understand each algorithm's habits and methods to correctly interpret the results. A strong collaboration between AI development engineers and clinical healthcare professionals will become even more critical in the future to apply AI in medicine.

References

- Fukuoka M, Yano S, Giaccone G, et al. Multi-institutional randomized phase II trial of gefitinib for previously treated patients with advanced non-small-cell lung cancer (The IDEAL 1 Trial) [corrected]. *J Clin Oncol.* 2003;21:2237–46.
- Morita S, Okamoto I, Kobayashi K, et al. Combined survival analysis of prospective clinical trials of Gefitinib for non–small cell lung Cancer with EGFR mutations. *Clin Cancer Res.* 2009;15:4493–8.
- Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. *Nat Methods.* 2010;7:248–9.
- Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods.* 2010;7:575–6.
- Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, Liu X. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet.* 2015;24:2125–37.
- Tang H, Thomas PD. Tools for predicting the functional impact of nonsynonymous genetic variation. *Genetics.* 2016;203:635–47.
- González-Pérez A, López-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score. *Condel Am J Hum Genet.* 2011; <https://doi.org/10.1016/j.ajhg.2011.03.004>.
- Ioannidis NM, Rothstein JH, Pejaver V, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet.* 2016;99: 877–85.
- Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD. ClinPred: prediction tool to identify disease-relevant nonsynonymous single-nucleotide variants. *Am J Hum Genet.* 2018;103:474–83.
- Oza AM, DiStefano MT, Hemphill SE, et al. Expert specification of the ACMG/AMP variant interpretation guidelines for genetic hearing loss. *Hum Mutat.* 2018;39:1593–613.
- Anderson D, Baynam G, Blackwell JM, Lassmann T. Personalised analytics for rare disease diagnostics. *Nat Commun.* 2019;10:5274.
- Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* 2019;47:D886–94.
- Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics.* 2015;31:761–3.
- Sundaram L, Gao H, Padigepati SR, et al. Predicting the clinical impact of human mutation with deep neural networks. *Nat Genet.* 2018;50:1161–70.
- Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human

- nonsynonymous and splice-site SNVs. *Genome Med.* 2020;12:103.
16. Hopf TA, Ingraham JB, Poelwijk FJ, Schärfé CPI, Springer M, Sander C, Marks DS. Mutation effects predicted from sequence co-variation. *Nat Biotechnol.* 2017;35:128–35.
 17. Riesselman AJ, Ingraham JB, Marks DS. Deep generative models of genetic variation capture the effects of mutations. *Nat Methods.* 2018;15:816–22.
 18. Sinai S, Kelsic E, Church GM, Nowak MA. Variational auto-encoding of protein sequences. *arXiv [q-bio.QM].* 2017.
 19. Ding X, Zou Z, Brooks CL III. Deciphering protein evolution and fitness landscapes with latent space models. *Nat Commun.* 2019;10:5644.
 20. McGee F, Novinger Q, Levy RM, Carnevale V, Haldane A. Generative capacity of probabilistic protein sequence models. *arXiv [cs.LG].* 2020.
 21. Gupta A, Zou J. Feedback GAN for DNA optimizes protein functions. *Nat Machine Intelligence.* 2019;1: 105–11.
 22. Wan C, Jones DT. Protein function prediction is improved by creating synthetic feature samples with generative adversarial networks. *Nat Machine Intelligence.* 2020;2:540–50.
 23. Zhang F, Lupski JR. Non-coding genetic variants in human disease. *Hum Mol Genet.* 2015;24:R102–10.
 24. Pena LDM, Jiang Y-H, Schoch K, et al. Looking beyond the exome: a phenotype-first approach to molecular diagnostic resolution in rare and undiagnosed diseases. *Genet Med.* 2018;20:464–9.
 25. Short PJ, McRae JF, Gallone G, et al. De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature.* 2018;555:611–6.
 26. Brandler WM, Antaki D, Gujral M, et al. Paternally inherited cis-regulatory structural variants are associated with autism. *Science.* 2018;360:327–31.
 27. Soemedi R, Cygan KJ, Rhine CL, Wang J, Bulacan C, Yang J, Bayrak-Toydemir P, McDonald J, Fairbrother WG. Pathogenic variants that alter protein code often disrupt splicing. *Nat Genet.* 2017;49:848–55.
 28. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, et al. Predicting splicing from primary sequence with deep learning. *Cell.* 2019;176: 535–548.e24.
 29. Zhou J, Troyanskaya OG. Predicting effects of non-coding variants with deep learning-based sequence model. *Nat Methods.* 2015;12:931–4.
 30. Zhou J, Park CY, Theesfeld CL, et al. Whole-genome deep-learning analysis identifies contribution of non-coding mutations to autism risk. *Nat Genet.* 2019;51: 973–80.
 31. Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat Genet.* 2018;50:1171–9.
 32. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med.* 2015;17:405–24.
 33. Griffith M, Spies NC, Krysiak K, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat Genet.* 2017;49:170–4.
 34. Lever J, Jones MR, Danos AM, Krysiak K, Bonakdar M, Grewal JK, Culibrk L, Griffith OL, Griffith M, Jones SJM. Text-mining clinically relevant cancer biomarkers for curation into the CIViC database. *Genome Med.* 2019;11:78.
 35. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv [cs.CL].* 2018.
 36. Beltagy I, Lo K, Cohan A. SciBERT: a Pretrained language model for scientific text. *arXiv [cs.CL].* 2019.
 37. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *arXiv [cs.CL].* 2019.
 38. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H. Domain-specific language model pretraining for biomedical natural language processing. *arXiv [cs.CL].* 2020.
 39. Gurovich Y, Hanani Y, Bar O, et al. Identifying facial phenotypes of genetic disorders using deep learning. *Nat Med.* 2019;25:60–4.
 40. Hsieh T-C, Mensah MA, Pantel JT, et al. PEDIA: prioritization of exome data by image analysis. *Genet Med.* 2019;21:2807–14.
 41. Bastarache L, Hughey JJ, Hebbings S, et al. Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. *Science.* 2018;359:1233–9.
 42. Martins Conde P, Sauter T, Nguyen T-P. An efficient machine learning-based approach for screening individuals at risk of hereditary haemochromatosis. *Sci Rep.* 2020;10:20613.
 43. Privé F, Aschard H, Blum MGB. Efficient implementation of penalized regression for genetic risk prediction. *Genetics.* 2019;212:65–74.
 44. Lello L, Raben TG, Yong SY, Tellier LCAM, Hsu SDH. Genomic prediction of 16 complex disease risks including heart attack, diabetes, breast and prostate Cancer. *Sci Rep.* 2019;9:15286.
 45. Qian J, Tanigawa Y, Du W, Aguirre M, Chang C, Tibshirani R, Rivas MA, Hastie T. A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK biobank. *PLoS Genet.* 2020;16:e1009141.
 46. Yang W, Soares J, Greninger P, et al. Genomics of drug sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* 2013;41:D955–61.
 47. Barretina J, Caponigro G, Stransky N, et al. The Cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature.* 2012;483:603–7.
 48. Li M, Wang Y, Zheng R, Shi X, Li Y, Wu F, Wang J. DeepDSC: a deep learning method to predict drug sensitivity of Cancer cell lines. *IEEE/ACM Trans*

- Comput Biol Bioinform. 2019. <https://doi.org/10.1109/TCBB.2019.2919581>.
49. Sharifi-Noghabi H, Zolotareva O, Collins CC, Ester M. MOLI: multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics. 2019;35:i501–9.
50. Chiu Y-C, Chen H-IH, Zhang T, Zhang S, Gorthi A, Wang L-J, Huang Y, Chen Y. Predicting drug response of tumors from integrated genomic profiles by deep neural networks. BMC Med Genomics 2019;12 Suppl 1:18, <https://doi.org/10.1186/s12920-018-0460-9>.
51. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, Staudt LM. Toward a shared vision for cancer genomic data. N Engl J Med. 2016;375:1109–12.
52. Costello JC, Heiser LM, Georgii E, et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nat Biotechnol. 2014;32:1202–12.
53. Chang Y, Park H, Yang H-J, Lee S, Lee K-Y, Kim TS, Jung J, Shin J-M. Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from Cancer genomic signature. Sci Rep. 2018;8:8857.
54. Preuer K, Lewis RPI, Hochreiter S, Bender A, Bulusu KC, Klambauer G. DeepSynergy: predicting anti-cancer drug synergy with deep learning. Bioinformatics. 2018;34:1538–46.
55. O’Neil J, Benita Y, Feldman I, et al. An unbiased oncology compound screen to identify novel combination strategies. Mol Cancer Ther. 2016;15:1155–62.
56. Xia F, Shukla M, Brettin T, et al. Predicting tumor cell line response to drug pairs with deep learning. BMC Bioinformatics. 2018;19:486.
57. Holbeck SL, Camalier R, Crowell JA, et al. The National Cancer Institute ALMANAC: a comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activity. Cancer Res. 2017;77:3564–76.
58. Jiang P, Huang S, Fu Z, Sun Z, Lakowski TM, Hu P. Deep graph embedding for prioritizing synergistic anticancer drug combinations. Comput Struct Biotechnol J. 2020;18:427–38.
59. Lumaka A, Cosemans N, Lulebo Mampasi A, et al. Facial dysmorphism is influenced by ethnic background of the patient and of the evaluator. Clin Genet. 2017;92:166–71.
60. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019;51:584–91.
61. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015;12:e1001779.
62. Sankar PL, Parker LS. The precision medicine Initiative’s all of us research program: an agenda for research on its ethical, legal, and social issues. Genet Med. 2017;19:743–50.
63. Birney E, Vamathevan J, Goodhand P. Genomics in healthcare: GA4GH looks to 2022. Cold Spring Harb Lab. 2017;203554.
64. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 618–26.
65. Mittelstadt B, Russell C, Wachter S. Explaining explanations in AI. In: Proceedings of the conference on fairness, accountability, and transparency. New York: Association for Computing Machinery; 2019. p. 279–88.
66. Tanaka Y, Tamada Y, Ikeguchi M, Yamashita F, Okuno Y. System-based differential gene network analysis for characterizing a sample-specific subnetwork. Biomol Ther. 2020. <https://doi.org/10.3390/biom10020306>.



Stem Cell Progression for Transplantation

79

Insights and Advances of Artificial Intelligence

Nazneen Pathan, Sharayu Govardhane, and Pravin Shende

Contents

Introduction	1098
Implementation of AI in Stem Cell Progression	1100
Machine Learning	1100
Computational Methods	1101
Deep Learning	1104
Conclusions	1105
References	1106

Abstract

Stem cell therapy displays various applications in biomedical fields in transplantation and treatment of neurodegenerative, cardiovascular, and cancerous diseases. The current understanding of the development of artificial intelligence (AI) contraption such as machine learning and deep learning is useful in many of the healthcare challenges like limited accuracy, detection, and identifying the nature of biological molecules. This review article focuses on the insights of applications of AI in the nature of stem cells and regenerative medicine, recognition of stem cell earlier to differentiation, characterization of stem cells employing

mathematical models, and estimation of toxicity related to stem cell transplantation. The reduction in complexity and costs by use of AI techniques for enhancement of clinical trials for stem cell and gene therapies by treating patients with detailed planning, forecasting clinical results, and maintaining the patients, applying to new outcomes from input data. The application and modern approach through digitalization by use of AI system for stem cell therapy and regenerative medicine play as an essential tool in highlighting the potential for improvement in fields of medicines in the future.

Keywords

Artificial intelligence · Stem cells · Transplantation · Machine learning

N. Pathan · S. Govardhane · P. Shende (✉)
Shobhaben Pratapbhai Patel School of Pharmacy and
Technology Management, SVKM'S NMIMS, Vile Parle
(West), Mumbai, India
e-mail: Pravin.Shende@nmims.edu

Abbreviation	
AI	Artificial intelligence
CBMIA	Content-based microscopic image analysis
CFD	Computational fluid dynamics
CLAD	Chronic lung allograft dysfunction
CNN	Convolutional neural networks
DL	Deep learning
DMARDs	Disease-modifying antirheumatic drugs
ESCs	Embryonic stem cells
FACS	Fluorescence-activated cell sorting
HCT	Hematopoietic cell transplantation
iPSCs	Induced pluripotent stem cells
ISNC	Improved supervised normalized cuts
ML	Machine learning
NN	Neural networks
RK4	4th-order Runge-Kutta
SC	Stem cells
SSD	Single Shot MultiBox Detector

Introduction

Artificial intelligence (AI) is the capability of a system to interpret data, acquire the learnings from data, and use the well-read data to complete the specific goals using flexible adaptation [1]. It is a permutation of deep learning (DL) and neural networks (NN) wherein the computational power mimics the functioning of the human brain by simplifying complex tasks. DL and computer vision, also known as Fourth Industrial Revolution Technologies, are key elements for the upcoming modern world [2]. Some of the convolutional neural networks (CNN) in medical and healthcare for diagnosis and treatment were studied for comparing human and AI performances and suggested “man with machine” approach [3]. AI shows a wide range of applications in health care and medicines like hematology, neurology, cell biology, oncology, cell therapy, and ophthalmology. AI tools also

reported improvement and development in hematopoietic cell transplantation (HCT) for pre-transplantation, post-transplantation, and grafting. Similarly, creating algorithms via AI to predict the risks involved in developing adverse events and assist clinicians in making healthier decisions leading to improved patient’s quality of life [1, 4]. AI displays ability to collect large data from healthcare systems, known as “big data” with help of different data bases like National Health and Nutrition Examination Survey Database, Genome Wide Association Studies, United Network for Organ Sharing, etc. AI provides risk assessment, general prognosis, correlating risk parameters with the disease condition, as well as clinical intervention strategies [5]. The application of AI for analysis of pluripotent stem cell (PSC)-derived colonies for overcoming the challenges like a time-consuming process, errors, and training-dependent that occur while manual identification leads to enhanced potential with greater accuracy and minimum errors in biomedical field. Stem cells (SC) show the ability of self-renewal division and produce two daughter cells wherein one cell undergoes final differentiation and the other one acquires the self-renewal process. The main types of stem cells are embryonic stem cells (ESCs), induced pluripotent stem cells (iPSCs), and adult stem cells (ASCs) where ESCs display the ability to differentiate into ectoderm, endoderm, and mesoderm germ cell layers like hematopoietic stem cells (HSCs), mesenchymal stem cells (MSCs), and neural stem cells (NSCs) [6]. The current technologies using AI serve immunohistochemistry and cytogenetics by phenotypically and genotypically characterizing plasma cells. Further, microfluid-assisted single-cell technology helped in the innovation of cancer stem cell subclonal evolution by developing the spatial distribution of quantitative subclonal measurement of multiple myeloma via AI-Med algorithms. Therefore, artificial intelligence-based integration of genome, epigenome, and pathological measurements are used for targeted therapy in the treatment of cancers (Fig. 1) [7]. Also, the molecular and cellular level tissue dynamics are easily obtained by using a three-dimensional computer model for developing and analyzing the

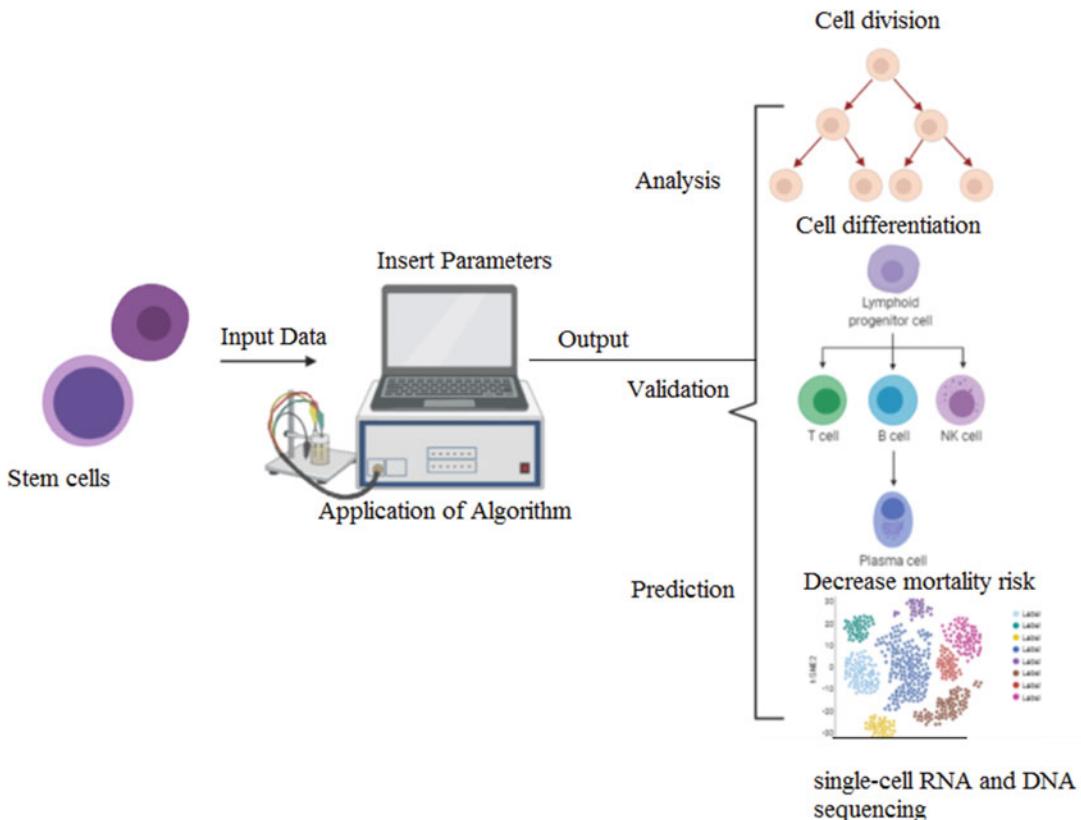


Fig. 1 AI approaches used in prediction, validation, and analysis of stem cells

hypothesis on mutations in the cells, identifying competitor stem cell and signaling pathways while activating intestinal tumor genesis. The multi-scale in silico computational model (Notch and Wnt signaling) serves as a tool for optimizing study designs used in tissue regeneration and tissue dynamics in the gut [8]. Similarly, the mismatching of the donor and acceptor specimen in transplantation leads to high level of acceptor rejection. The replacement of bronchoscopy with AI-driven techniques on chronic lung allograft dysfunction (CLAD) helps in forecasting immune-mediated allograft damage and optimizing reproducible better outcomes after lung transplantation by pulmonary diagnosis serving as next-generation tool for converting paradigm from “Detect” to “Detect, Quantify, and Predict” [9]. Stem cells display wide range of applications in treatment of various conditions such as diabetes, vision impairment, infertility, HIV,

Parkinsonism, osteoarthritis, etc. Currently stem cell therapy also showed potential in treatment of SARS-CoV-2-related diseases like pneumonia and other lung injuries due to virus leading to significant impact to save life of patient [10, 11]. Also, certain mesenchymal cells are used in organ transplantation for auto-immune disorders. The umbilical cord mesenchymal cells in combination with disease-modifying antirheumatic drugs (DMARDs) showed effective treatment in patients suffering with rheumatoid joint inflammation [12]. Further, the regulation of cultured stem cells by mechanical signals and bioactive agents displays liposomal-based stem cell therapy for multimodal cancer treatment [13]. The multidisciplinary computational immunobiology-based predictive models disclose mechanisms and pathways for organ transplantation through proteomics, genomics, protein modification, analysis of gene expression,

development of statistical models, identification of gene networks, host-pathogen interactions, and bioinformatics. The applications of such high-end multi-scale models in facilitating immune response and other cell-related discrete events like molecular interactions and Van der Waal forces in the proteins provide a next-generation tool for antibody editing, tissue regeneration, and peripheral immune regulation [10, 14]. Bioinformatics analyses and functional validation, a type of AI neural network, defined the transcriptomes of progenitors present at various stages of embryo such as erythroid-, megakaryocyte-, and leukocyte-committed progenitors, as well progenitor markers and membrane markers like CD44, CD326, ICAM2/CD9, and CD18. Also, single-cell RNA sequencing with protein expression analysis and CITE sequencing helped in identifying transcriptional factors and molecular pathways of in vitro-generated HSC-like cells from the fetal liver and comparing in vitro-generated progenitors with in vivo-produced cells, exploring improvement in the production of SC in the future [15].

Implementation of AI in Stem Cell Progression

Machine Learning

Machine learning is a branch of AI where the algorithms are automatically learned by the models on basis of examples and predict the response to a treatment constructed on different parameters related to the patients, drug, biological cells, etc., overtaking regression methods and attaining accuracies with enhanced and easy interpretation of obtained results [16]. The combination of AI and ML aids in problem-solving, indulging different languages, distinguishing various images, voices, and additional “smart” tasks. The flexibility of machine learning is better than the typical statistical method and is better suited for predictive problems providing advantageous technology for healthcare systems [13, 14]. The different algorithms of machine learning are applied in the determination of cell morphology, graft transplant, and molecular imaging cardiac

tissue contractility. Moreover, ML shows a wide range of applications in the systemic analysis of the method, optimization, and diagnosis of disease conditions with higher accuracy and precision leading to better medical decision-making [11, 12]. The three types of machine learning model mainly used are 1) supervised learning where the predictions are revealed by labeled data set, 2) unsupervised learning wherein the patterns and clusters are identified by the algorithm, and 3) reinforcement learning where the sequences are detected on basis of trial error longitudinal from data. These models are applied for prediction of survival of allogeneic stem cell transplantation, gene expression profiles, and administration of vasopressor for critically ill patients [17]. New screening techniques were developed by utilizing machine learning for the improvement of recapitulation of human cardiac functioning and responses to drug using human pluripotent stem cells. Further, the comprehensive analysis of 17 parameters and contractility on human pluripotent stem cells-cardiac tissue strips exposed to drugs like norepinephrine, ramipril, flecainide, and lisinopril was measured by automated drug classification for predicting the mechanism of anonymous cardioactive drugs [18]. The machine learning-based analysis of stem cell donor availability associated with age, gender, and other characteristics was established for predicting the potential donors employing the receiver-donor curve to reduce the time for transplantation. The National Bone Marrow dataset was separated into two: training accuracy set and testing accuracy set using domain-specific knowledge [19]. A high risk in prediction mortality associated with hematopoietic stem cell transplantation is observed, and the use of the ML in silico computational method helped in identifying the HCT-comorbidity index. The optimized ML model was developed with a type of algorithm, data sets, and number of variables that resulted in 6 ML algorithm models including 24 variables filtered down to minimum variables and compared accordingly to the AUCs. The prediction of morbidity via ML technology showed more efficient results as it overcomes the limitations like lengthy methodology, medical procedures,

and cost [20]. Also, the mortality rate of 55% in patients was due to sepsis during hematopoietic stem cell transplantation. Therefore, the prediction of sepsis was achieved 41.5 h beforehand on 313 patients using ML survival random forest model to prevent the patient suffering at the time of engraftment admission for autologous and allogeneic HCT. The different variables like patient's blood calcium, ferritin, creatinine level, glucose, lymphocyte percentage, WBC count, heart rate, platelet, body temperature, blood pressure, and respiratory rate were observed, and differentiating patterns between septic and the non-septic patient were determined to prevent organ damage due to sepsis [21]. Another ML model is known as one-class logistic regression to remove the transcriptomic and epigenetic specific set from PSC and predict biological mechanisms from oncogenic cells in the tumor microenvironment. Novel targeted therapies developed from the stemness prediction by revealing data molecular heterogeneity. The results showed DNA demethylation stemness index, correlation of mRNAs and mDNAs, and cancerous undifferentiated cells further helping in mutations in genes representing oncogenicity and analyzing the stem cell-based immunogenic therapy for the safety of the patient [22]. A supervised ML system based on SimFSC software in combination with voltage-sensitive dyes for classifying the chronotropic drug (propranolol and isoproterenol) effects on cardiomyocyte by using algorithm, resulted in predicting the membrane polarization occurring due to drugs on the cardiomyocyte [23]. Single-cell molecular-based technologies established a method to provide HSC profile model wherein the single-cell transcriptomes were detected from the murine bone marrow by gene expressions. The highly variable genes through protein-coding were divided into positive and negative significant correlation data sets known as "Molo" and "Nomo," respectively. This data represented the functionality and transcriptional changes and genetic perturbations and provided new insights into disrupted hematopoietic situations [24]. The unsupervised type of ML technique known as K-means++ was used in cell-based therapy for quantification of magnetic

particle imaging. The application of canonical algorithm to in vitro (VivoTrax-labeled islets), in vivo (labeled islets under the kidney capsule), and ex vivo (excised kidney grafts) data using islets and mouse models for segmenting the regions of interest in formed images and quantifying iron (0.5–7.5 µg). The 3D phantom imaging and optimization of the algorithm with intra-class correlation coefficient and inter-rater reliability validation followed by statistical analysis (ANOVA) showed higher accuracy in low time. Further, this methodology is also applied to iPSC or chimeric antigen receptor T cells for analysis, magnetic imaging, quantifying uptake of nanoparticle monitoring post-transplant graft loss, fluctuation of transplant cells, etc. and serving potential in clinically transplantation therapies for type 1 diabetes [25].

Computational Methods

The interconnected molecular network in the cells is known as "systems biology." The assessment of high throughput of biological complexes, molecular-level interactions, DNA variants, gene-level modification, and recognizing the role of biomolecules is feasible via computational modeling and simulation models (Fig. 2). The combination of computation modeling-based AI in cancer, diabetes, cardiology, and neurology helps in better treatment of disease and understanding of health care [26]. A computational cell kinetics model for understanding the intercellular signaling networks and alteration of lineage specification in early progenitor cell culture population also known as HSC lineage engineering was used in a study on the femur and tibia bone marrow of the female C57BL/6 mice, wherein STELLA and Berkeley Madonna computational models were used to determine the high-impact parameters useful in cell expansion and transplantation. The cell kinetics were observed on four types of cell-secreted biomolecules such as differentiation stimulators and proliferation stimulators and inhibitors for identifying cell differentiation rates and plan experimental designs [27]. Similarly, a study demonstrated a full automatic computer-aided

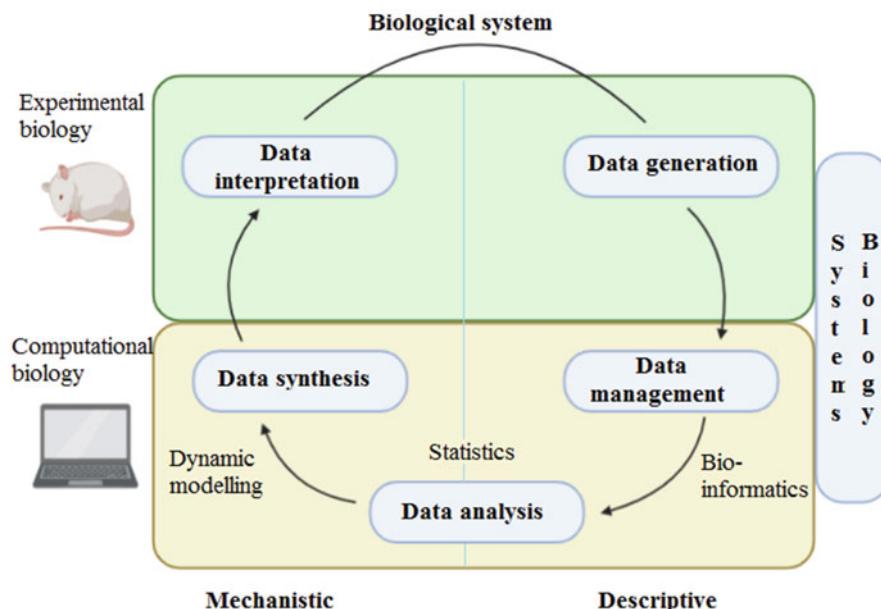


Fig. 2 Computational approaches in the biological system

clustering system for finding relations between cell differentiation and morphological and structural changes in phenotype. A computer model, known as a multi-stage Content-based Microscopic Image Analysis (CBMIA), was applied for segmentation of images, extraction, selection, and fusion techniques. The original microscopical image of multiple stem cells was introduced to Improved Supervised Normalized Cuts (ISNC) segmentation algorithm leading to higher similarity (92.4%), enhanced specificity (96.0%), and greater accuracy (107.8%). The three algorithms known as normalized cuts, supervised normalized cuts, and improved supervised normalized cuts were functionalized, and Hu and Zernike moments (most robust characteristics) are selected and applied to the k-means clustering method revealing the enhancement in the cell description on phenotypic level of the stem cells [28]. Another study displayed the use of the computational method and mathematical modeling approaches to defining underlying processes and the fate of ESC in mouse cells. The three transcriptional factors, namely, Oct4, Sox2, and homeobox protein NANOG were used to measure the gene expression in human and mouse ESC. The identification of network components and topology by

descriptive, dynamic modeling, and explorative statistics for the analysis of gene expression was carried out by using Boolean network approaches and Bayesian network model. Different algorithms like gene regulatory network, K-means clustering (clustering algorithm), multi-scale model (mathematical framework), ordinary differential equations (mathematical equation), principle component analysis (statistical method), support vector machine, stochastic differential equations, etc. are applied for identification of topology, several genes, changes occurring in genes and cells, and behavior variability and interactions [29]. According to a study, the migration of hematopoietic stem cells inside the bone marrow is proposed by the use of a mathematical model where in vitro HSCs were simulated with a finite element tool Gascoigne resulting in improvement of the solution of an accurate chemotaxis model with nonlinear boundary situations. The HSCs formed in the bone marrow migrate to the blood periphery and differentiate; therefore the analysis and study of transplantation would be helpful for issues such as forecasting the consequences of various treatments for specific blood-related disorders and limiting the time where the patient is missing their effective

immune system [30]. A multi-parameter immunophenotyping model identified differentiation of HSC in healthy donors ($n = 10$) and determined plausible hierarchies by quantitative model selection to confirm robustness of model selection at varying levels of noise. HSC differentiation is initiated by progenitor compartments of reducing plasticity and enhancing maturity in a hierarchical order. Seven cell-type compartments were determined via computational modeling wherein there was 90% differentiation of multipotent lymphoid progenitor into granulocyte-monocyte progenitors that were plausible. The established data of ten HSC donors were analyzed by in vitro fluorescence-activated cell sorting (FACS) process followed by Bayesian information criterion selection [31]. Stem cell therapy also showed effective treatment in HIV infections wherein, the mathematical equation-based system known as 4th-order Runge-Kutta (RK4) was applied to obtain numerical solutions graphically and transfer model to provide an enhancement in the detection and identification of CD4⁺T cells. The relationship between the SC for production of T cells and population of CD4⁺T was determined by symmetric and asymmetric self-renewal cells and differentiated cells with probability factor eventually providing global stability at the prevalent point [32]. Further, the computer-aided classification from stem cell fate analysis by quantified 2D and 3D images and improvement of accuracy and recognition via optimizing the confocal laser scanning microscope was carried out by classification protocol known as Volocity® cytoplasmic cell marker. The validated and analyzed images of hNSC progeny showed quicker (twofold) and higher strength than human-based image analysis. The parameters like object size, the intensity of color, and z-step interval are applied for quantification of different human cell lineages in combination with microscope image capture systems [33]. The determination of HSC biology in transcose-6-phosphate dehydrogenase heterozygous cats was demonstrated by using computational-simulated gene therapy through analysis of autologous transplantation studies where, the Splus statistical programming language approach with

parameters like the intensity of replication, apoptosis, and differentiation was used wherein the marked clones controlled the hematopoiesis process. Therefore, the simulated computer model showed applications in hematopoietic cell-related disorders such as sickle cell anemia, thalassemia, and myeloproliferative disorders like myelogenous leukemia, etc. with a 10% survival advantage [34]. A hydrodynamic environment in a bioreactor was described by developing a computational fluid dynamic (CFD) model wherein the determination of ESC size was correlated with eddy size. The hydrodynamics of suspension bioreactors affects cell viability, proliferation, differentiation, and pluripotency. The cell showed 95% gene expression of SSEA-1, Sox-2, and NANOG without changing the distribution pattern in various bioreactor parameters. Moreover, decrease in eddy size enhances the potential cell damage with a higher volume shear rate (40 times) thereby affecting the morphological, structural, phenotypic, and functional abilities of the stem cells. Therefore, CFD is applied in the production of large-scale stem cell culture processes [35]. A similar study conducted by using CFD on seven different scale-up parameters was used to analyze the agitation rate of the bioreactor. The model identified a constant maximum shear stress/dissipation rate required to attain the murine ESC cultured in aggregate sizes and therefore effective stem cell growth scale-up was achieved without high agitation rate by CFD [36]. A study demonstrated live camera-based stem cell tracking system for identification of variables in donor HSC (human umbilical cord mesenchymal stem cells). The improvement in cellular activity was achieved by applying automated high-throughput time-lapsed microscopic imaging, with custom-built live automated cell imager and microscope attached to modern digital camera. Further, an environmental control system with robotic microscope stage provides parameters to be measured via time-lapsed images like cell morphology, proliferation, velocity, and fluorescence leading to analysis of bone morphogenic protein for assessing osteogenic potential by measuring the level of alkaline phosphatase in populations [37]. The analysis of motility of spermatozoa in

the testis of genetically sterile mice after spermatogonial stem cell transplantation from testicular cells of fertile donor mice by computer-assisted validation helped in determination of the sperm motility analysis technology, quality of sperm flagella motion, and three velocities such as curvilinear path, straight path, and average path. Thus, the cause for reduction of fertilization rate post-IVF was established by using computer and AI models wherein the decreased number of motile spermatozoa and the earlier reduction of the individual sperm movement characters led to increase in fertilization [38]. The HSC transplantation from the bone marrow to peripheral blood is also used for the treatment of chronic myeloid leukemia. There exists already available IT applications for the donor search process such as the German Marrow Donor Information System (GERMIS), Donor-Center's Communication Workbench (DoCCom), Search Coordinator's Communication Workbench (SeCCom), etc. used for donor management, patient administration, transplant request management, and handling communication (telematics) online [38].

Deep Learning

Deep learning is a type of machine learning wherein the neural networks mimic the neuronal network present in the human brain. The neural network involves several layers of neurons and translates input signal into output signal also known as the predictive value. The various applications of DL are clinical studies, cell imaging, genomic analyses, protein prediction, identification of end-stage hematopoietic lineage from microscope images, etc. [39]. A quality control study based on deep learning was performed on human pluripotent stem cells-derived cardiomyocytes employing CNN wherein the CNN was able to discriminate the abnormal cells within normal cell culture with mean force of 0.89 and speed of 2000 images per second. The CNN works better in ethics, speed, cost, and labor with improved accuracy, correctness, and efficient performance when associated to humans [40]. The knowledge of the topological structure and surface morphology of human bone marrow stromal

stem cells via DL neural network and laser topography displayed tissue fabrication, tissue structure, and cell signaling useful for predicting structural functionality, cell response, and cell behavior. The DL statistical probability tool developed 203 fluorescent images that assisted in revealing variations between donor cells like proliferation variation and percentage of non-responders to osteogenic differentiation and also enhanced the predictive capability of the model unlike photolithography that is time-consuming and costly [41]. Another DL-based cell motion analysis, quantitative model for identification and evaluation of cultured keratinocyte stem cell via the non-invasion method was studied in keratinocytes to address alterations. Further, Single Shot MultiBox Detector (SSD) deep learning model was applied to produce phase-contrast images of human epidermal keratinocytes wherein the velocity vector of single keratinocytes cells was observed by the quantitative estimation of behavior differences of keratinocyte by varying culture conditions. Also, automated cell tracking and clonal analysis showed unique patterns of motion leading to the identification of the stem cell colonies with velocity information of cells and without labeling of human cells [42]. Similarly, a U-Net-based algorithm was used to measure the various types of parameters affecting the formation of cells, hiPSC colony, single cells, dead cells, and differentiated cells and other parameters like shape, size, center of gravity attachments of other cells, etc. The automated cell analysis enables an average recording time of less than 1 min wherein the image provided displayed a total size of about 610 MP and each pixel denoting $1.69\mu\text{m}$. This method served hiPSCs for biomedicine applications in transplantation by assisting automated generation and cultivating high-quality hiPSCs [43]. Also, the identification of endothelial cells obtained from hiPSCs by utilizing the K-fold cross-validation based on four independent datasets was carried out by comparing with immunofluorescence staining for CD31 (marker). The blocks were inserted using phase-contrast images, and the input data was validated by network optimization using CNN leading to error analysis and improvement in performance and efficient formation of

endothelial cells (20%–35%). The data adjustment was carried out by input number of blocks (labeled and unlabeled cells), size of blocks, and fixing the ratio of targeted blocks followed by LeNet and AlexNet [44]. Furthermore, enhancement in accuracy up to 99% was observed in less than 20 min on using DL neural networks for the identification of differentiated cells from undifferentiated cells. Precision, recall, and F1 value of two CNN models (ResNet50-SA and DenseNet50-SA) were used wherein the differentiated and pluripotent stem cells were determined at various stages of the cell cycle. The CNN was also able to differentiate serum and leukemia inhibitory factor from the ESC wherein the primed biological markers FGF5, Oct6, Dnmt3A, and Otx2 were upregulated and pluripotency markers Klf4, Nanog, Esrrb, and Tbx3 were downregulated at the end of 48 h. The optimization of the algorithm was developed by using epoch also known as learning rate at each training cycle [45]. Another study reported the recognition of acute cellular rejection during lung transplant biopsies using digital pathology and AI technology. After 3349 annotations, the vascular component of acute cellular rejection was differentiated from normal alveolar lung tissue with high accuracy (95%). Therefore, the DL AI algorithm is also practiced in diagnosis and the detection of cellular rejection during transplantation [46]. Similarly, the assessment of cardiac allograft vasculopathy and progression conventionally performed by 3D OCT imaging and wall

layer quantification analysis was carried out by novel DL method for atherosclerosis. The DL CNN leave-20%-out approach was evaluated in OCT pullbacks assimilated at 1 and 12 months after heart transplantation in 116 patients wherein DL method displayed 86.3% correctness in images within 10 s when compared to manual method (1 h). The 3D bioprinting technique by using computer programs is used for gene expression in cancers, tissue engineering, as well as skin and vascular regeneration [47, 48, 49, 50] (Fig. 3).

Conclusions

The current article provides insights into various newer permutation methods in combination with stem cells to understand and mimic the mechanism of transplantation. The application of artificial intelligence with stem cell therapy provides the advantage for optimization of clinical trials, development of precise trajectories for transplantation, reduction of error, application of up-to-date technologies, and prediction of results. Integration of artificial intelligence with stem cells is foreseeable; however it is still considered a deficient technique in comparison with the conventional tools. With the growth in regenerative therapy in transplantation, innovative techniques such as XplOit, allogeneic hematopoietic stem cell transplantation, chimeric antigen receptor (CAR) T-cell therapy, bioprint, etc. proved to be effective in combination with algorithm performance. With

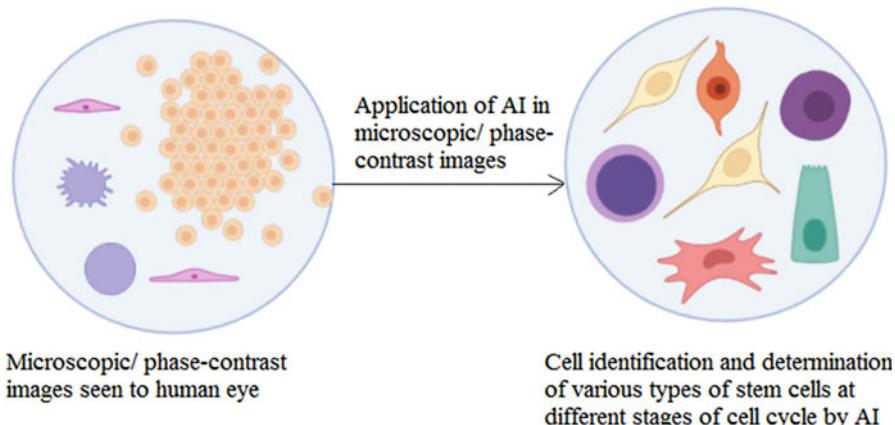


Fig. 3 Difference between microscopic images seen to human eye and via using AI

the application of the algorithm technique, stem cell therapy will be possibly able to modify the conventional transplantation therapies. Although the application of artificial intelligence in the field of stem cell transplantation acts as an alternative for the risky conventional method, predictive analysis remains a concern. Though the application of stem cells has helped to overcome such obstacles, the efficiency and safety regarding the transplantation with the help of artificial intelligence will be resolute in the near future.

References

1. Alsuliman T, Humaidan D, Sliman L. Machine learning and artificial intelligence in the service of medicine: necessity or potentiality? *Curr Res Transl Med* [Internet]. 2020;68(4):245–51. <https://doi.org/10.1016/j.retram.2020.01.002>.
2. Kakani V, Nguyen VH, Kumar BP, Kim H, Pasupuleti VR. A critical review on computer vision and artificial intelligence in food industry. *J Agric Food Res* [Internet]. 2020;2:100033. <https://doi.org/10.1016/j.jafr.2020.100033>.
3. Hekler A, Utikal JS, Enk AH, Hauschild A, Weichenthal M, Maron RC, et al. Superior skin cancer classification by the combination of human and artificial intelligence. *Eur J Cancer*. 2019;120:114–21.
4. Muhsen IN, Elhassan T, Hashmi SK. Artificial intelligence approaches in hematopoietic cell transplantation: a review of the current status and future directions. *Turkish J Hematol*. 2018;35(3):152–7.
5. Thongprayoon C, Kaewput W, Kovvuru K, Hansrivijit P, Kanduri SR, Bathini T, et al. Promises of Big Data and artificial intelligence in nephrology and transplantation. *J Clin Med*. 2020;9(4):1107.
6. Ramakrishna RR, Hamid ZA, Zaki WMDW, Huddin AB, Mathialagan R. Stem cell imaging through convolutional neural networks: current issues and future directions in artificial intelligence technology. *PeerJ*. 2020;8:e10346.
7. Lee LX, Li SC. Hunting down the dominating subclone of cancer stem cells as a potential new therapeutic target in multiple myeloma: an artificial intelligence perspective. *World J Stem Cells*. 2020;12(8):706–20.
8. Thalheim T, Buske P, Przybilla J, Rother K, Loeffler M, Galle J. Stem cell competition in the gut: insights from multi-scale computational modelling. *J R Soc Interface*. 2016;13(121):20160218.
9. Shigemura N. Transforming diagnostics in lung transplantation: from bronchoscopy to an artificial intelligence-driven approach. *Am J Respir Crit Care Med*. 2020;202(4):486–7.
10. Desai D, Shende P. Nanoconjugates-Based Stem Cell Therapy for the Management of COVID-19. *Stem Cell Rev Reports* 2020. <https://doi.org/10.1007/s12015-020-10079-6>.
11. Shende P, Bhandarkar S, Prabhakar B. Heat shock proteins and their protective roles in stem cell biology. *Stem Cell Rev Rep*. 2019;15(5):637–51.
12. Shende P, Rodrigues B, Gaud RS. Transplantation and alternatives to treat autoimmune diseases. *Adv Exp Med Biol*. 2018;1089:59–72.
13. Mandpe P, Prabhakar B, Shende P. Role of liposomes-based stem cell for multimodal cancer therapy. *Stem Cell Rev Rep*. 2020;16(1):103–17.
14. Vásquez-montoya GA, Danobeitia JS, Fernández LA, Hernández-ortiz JP. Computational immuno-biology for organ transplantation and regenerative medicine. *Transplant Rev* [Internet]. 2016. <https://doi.org/10.1016/j.tre.2016.05.002>.
15. Fidanza A, Stumpf PS, Ramachandran P, Tamagno S, Babtie A, Lopez-Yrigoyen M, et al. Single-cell analyses and machine learning define hematopoietic progenitor and HSC-like cells derived from human PSCs. *Blood*. 2020;136(25):2893–904.
16. Squarcina L, Villa FM, Nobile M, Grisan E, Brambilla P. Deep learning for the prediction of treatment response in depression. *J Affect Disord* [Internet]. 2021;281:618–22. <https://doi.org/10.1016/j.jad.2020.11.104>.
17. Shouval R, Fein JA, Savani B, Mohty M, Nagler A. Machine learning and artificial intelligence in haematology. *Br J Haematol*. 2021;192(2):239–50.
18. Lee EK, Tran DD, Keung W, Chan P, Wong G, Chan CW, et al. Machine learning of human pluripotent stem cell-derived engineered cardiac tissue contractility for automated drug classification. *Stem Cell Rep* [Internet]. 2017;9(5):1560–72. <https://doi.org/10.1016/j.stemcr.2017.09.008>.
19. Sivasankaran A, Williams E, Albrecht M, Switzer GE, Cherkassky V, Maiers M. Machine learning approach to predicting stem cell donor availability. *Biol Blood Marrow Transplant* [Internet]. 2018;24(12):2425–32. <https://doi.org/10.1016/j.bbmt.2018.07.035>.
20. Shouval R, Labopin M, Unger R, Giebel S, Ciceri F, Schmid C, et al. Predictive limitations of hematopoietic stem cell transplantation associated mortality: a machine learning in-silico analysis of the EBMT – acute leukemia working party registry. *Biol Blood Marrow Transplant*. 2015;21(2):S310–1.
21. Dadwal SS, Eftekhari Z, Thomas T, Munu J, Yang D, Mokhtari S, et al. A dynamic machine-learning based prediction model for sepsis in patients undergoing hematopoietic stem cell transplantation. *Biol Blood Marrow Transplant* [Internet]. 2018;24(3):S373–4. <https://doi.org/10.1016/j.bbmt.2017.12.457>.
22. Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, et al. Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell*. 2018;173(2):338–354.e15.
23. Heylman CM, Datta R, Conklin BR, George SC, Gratton E. Classifying the electrophysiological effects of chronotropic drugs on human induced pluripotent stem cell-derived cardiomyocytes using voltage

- sensitive dyes and supervised machine learning. *Biophys J* [Internet]. 2015;108(2):110a. <https://doi.org/10.1016/j.bpj.2014.11.624>.
- 24. Hamey FK, Göttgens B. Machine learning predicts putative hematopoietic stem cells within large single-cell transcriptomics data sets. *Exp Hematol*. 2019;78: 11–20.
 - 25. Hayat H, Sun A, Hayat H, Liu S, Talebloo N, Pinger C, et al. Artificial intelligence analysis of magnetic particle imaging for islet transplantation in a mouse model. *Mol Imaging Biol*. 2021;23(1):18–29.
 - 26. Winslow RL, Trayanova N, Geman D, Miller MI. Computational medicine: translating models to clinical care. *Sci Transl Med*. 2012;4(158):1–12.
 - 27. Mahadik B, Hannon B, Harley BAC. A computational model of feedback-mediated hematopoietic stem cell differentiation in vitro. *PLoS One*. 2019;14(3):1–21.
 - 28. Li C, Huang X, Jiang T, Xu N. ScienceDirect Full-automatic computer aided system for stem cell clustering using content-based microscopic image analysis. 2017;7.
 - 29. Herberg M, Roeder I. Computational modelling of embryonic stem-cell fate control. *Development*. 2015;142(13):2250–60.
 - 30. Bencheva G. Computer modelling of haematopoietic stem cells migration. *Comput Math Appl*. 2012;64: 337–49.
 - 31. Bast L, Buck C, Judith S, Katharina S. Computational modeling of stem and progenitor cell kinetics identifies plausible hematopoietic lineage hierarchies. *iScience* 24, 2021. <https://doi.org/10.1016/j.isci.2021.102120>.
 - 32. Alqudah MA, Aljahdaly NH. Global stability and numerical simulation of a mathematical model of stem cells therapy of HIV-1 infection. *Journal of computer science*. 2020;45. <https://doi.org/10.1016/j.jocs.2020.101176>.
 - 33. Piltti KM, Haus DL, Do E, Perez H, Anderson AJ, Cummings BJ. Computer-aided 2D and 3D quantification of human stem cell fate from in vitro samples using Velocity high performance image analysis software. *Stem Cell Res*. 2011;7:256–63.
 - 34. Abkowitz BJL, Catlin SN, Guttorp P. Strategies for hematopoietic stem cell gene therapy: insights from computer simulation studies. *Blood*. 1997;89(9): 3192–3198. <https://doi.org/10.1182/blood.V89.9.3192>.
 - 35. Borys ABS, Le A, Roberts EL, Rohanisarvestani L, Hsu CY, Wyma AA, et al. Using Computational Fluid Dynamics (CFD) Modeling to understand Murine Embryonic Stem Cell Aggregate Size and Pluripotency Distributions in Stirred Suspension Bioreactors. *Journal of Biotechnology*. S0168-1656(19)30811-9 2019. <https://doi.org/10.1016/j.jbiotec.2019.08.002B> IOTEC.
 - 36. Kallos S, Breanna S, Borys, Erin L, Roberts, An Le, Michael, Scale-up of Embryonic Stem Cell Aggregate Stirred Suspension Bioreactor Culture Enabled by Computational Fluid Dynamics Modeling. *Biochemical Engineering Journal*. 2018. <https://doi.org/10.1016/j.bej.2018.02.005>.
 - 37. Deasy BM, Chirieleison SM, Witt AM, Peyton MJ, Bissell TA. Tracking stem cell function with computers via live cell imaging: identifying donor variability in human stem cells. *Operative techniques in orthopaedics* [Internet]. 2010;20(2):127–35. <https://doi.org/10.1053/j.oto.2009.10.010>.
 - 38. Goossens E, De Block G, Tournaye H. Computer-assisted motility analysis of spermatozoa obtained after spermatogonial stem cell transplantation in the mouse. *Fertil Steril*. 2008;90:1411. <https://doi.org/10.1016/j.fertnstert.2007.08.035>.
 - 39. Kusumoto D, Yuasa S. The application of convolutional neural network to stem cell biology. *Inflamm Regen*. 2019;39(1):1–7. <https://doi.org/10.1186/s41232-019-0103-3>.
 - 40. Orita K, Sawada K, Koyama R, Ikegaya Y. Deep learning-based quality control of cultured human-induced pluripotent stem cell-derived cardiomyocytes. *J Pharmacol Sci* [Internet]. 2019;140(4):313–6. <https://doi.org/10.1016/j.jphs.2019.04.008>.
 - 41. Mackay BS, Praeger M, Grant-Jacob JA, Kanczler J, Eason RW, Oreffo ROC, et al. Modeling adult skeletal stem cell response to laser-machined topographies through deep learning. *Tissue and Cell* [Internet]. 2020;67:101442. <https://doi.org/10.1016/j.tice.2020.101442>.
 - 42. Nanba D, Hirose T, Toki F, Nishimura EK, Kotoku J. 593 Label-free identification of human keratinocyte stem cells by deep learning-based quantitative cell motion analysis. *J Invest Dermatol* [Internet]. 2019;139(9):S316. <https://doi.org/10.1016/j.jid.2019.07.597>.
 - 43. Piotrowski T, Rippel O, Elanzew A, Nießing B, Stucken S, Jung S, et al. Deep-learning-based multi-class segmentation for automated, non-invasive routine assessment of human pluripotent stem cell culture status. *Comput Biol Med*. 2021;129:104172.
 - 44. Kusumoto D, Lachmann M, Kunihiro T, Yuasa S, Kishino Y, Kimura M, et al. Automated deep learning-based system to identify endothelial cells derived from induced pluripotent stem cells. *Stem Cell Rep* [Internet]. 2018;10(6):1687–95. <https://doi.org/10.1016/j.stemcr.2018.04.007>.
 - 45. Waisman A, La Greca A, Möbbs AM, Scarafia MA, Santín Velazque NL, Neiman G, et al. Deep learning neural networks highly predict very early onset of pluripotent stem cell differentiation. *Stem Cell Rep*. 2019;12(4):845–59.
 - 46. Davis H, Glass C, Davis RC, Glass M, Pavlisko EN. Detecting acute cellular rejection in lung transplant biopsies by artificial intelligence: a novel deep learning approach. *J Heart Lung Transplant* [Internet]. 2020;39(4):S501–2. <https://doi.org/10.1016/j.healun.2020.01.100>.
 - 47. Shende P, Trivedi R. 3D printed bioconstructs: regenerative modulation for genetic expression. *Stem Cell Rev Rep*. 2021. <https://doi.org/10.1007/s12015-021-10120-2>.
 - 48. Mullan S, Chen Z, Pazdernik M, Zhang H, Wahle A, Melenovsky V, et al. Deep learning facilitates automation of wall layer quantification in heart

- transplant coronary OCT. *J Hear Lung Transplant*. 2019;38(4):S281. <https://doi.org/10.1016/j.healun.2019.01.702>.
49. Sivapalaratnam S. Artificial intelligence and machine learning in haematology. *Br J Haematol*. 2019;185(2): 207–8.
50. Senanayake S, White N, Graves N, Healy H, Baboolal K, Kularatna S. Machine learning in predicting graft failure following kidney transplantation: a systematic review of published predictive models. *Int J Med Inform* [Internet]. 2019;130:103957. <https://doi.org/10.1016/j.ijmedinf.2019.103957>.



Artificial Intelligence in Blood Transcriptomics

80

Stefanie Warnat-Herresthal, Marie Oestreich,
Joachim L. Schultze, and Matthias Becker

Contents

Introduction	1110
Blood Transcriptomics in Clinics: Methods, Features, Pitfalls	1112
Background on Artificial Intelligence in Biology	1113
Research Development Towards Clinical Applications	1119

S. Warnat-Herresthal

Systems Medicine, Deutsches Zentrum für
Neurodegenerative Erkrankungen (DZNE), Bonn,
Germany

Genomics and Immunoregulation, Life & Medical
Sciences (LIMES) Institute, University of Bonn, Bonn,
Germany

e-mail: stefanie.herresthal@uni-bonn.de

M. Oestreich · M. Becker (✉)

Systems Medicine, Deutsches Zentrum für
Neurodegenerative Erkrankungen (DZNE), Bonn,
Germany

e-mail: marie.oestreich@dzne.de;
matthias.becker@dzne.de

J. L. Schultze

Systems Medicine, Deutsches Zentrum für
Neurodegenerative Erkrankungen (DZNE), Bonn,
Germany

Genomics and Immunoregulation, Life & Medical
Sciences (LIMES) Institute, University of Bonn, Bonn,
Germany

PRECISE Platform for Single Cell Genomics and
Epigenomics at German Center for Neurodegenerative
Diseases (DZNE) and the University of Bonn, Bonn,
Germany

e-mail: joachim.schultze@dzne.de

© Springer Nature Switzerland AG 2022

N. Lidströmer, H. Ashrafiyan (eds.), *Artificial Intelligence in Medicine*,
https://doi.org/10.1007/978-3-030-64573-1_262

1109

Ethical Considerations, Data Security and Federated Learning	1119
Outlook	1120
Cross-References	1120
References	1121

Abstract

While the analysis of high-dimensional gene expression data by means of artificial intelligence (AI) has become an integral part of basic biomedical and immunological research, the great potential of this approach to substantially advance clinical diagnosis as well as medical routines has not become a reality yet. Significant advancements in both RNA-sequencing technologies and in the development of suitable algorithms for high-dimensional data, however, enabled a plethora of clinically oriented research that proposes promising concepts for systems-oriented and individualized medicine based on transcriptomic data. The blood transcriptome is of particular interest, as it is easily accessible and has proven to be a good predictor of blood and non-blood-based diseases thus making it a prime candidate to develop new powerful diagnostic approaches. The addition of concepts to the medical armamentarium, however, implicates ethical, legal, and technical aspects of data governance, which are a prerequisite to successfully integrate data-driven AI technologies into clinical routines.

Keywords

RNA-seq · Transcriptomics · Artificial intelligence · Machine learning · Diagnosis · Systems medicine

Introduction

The emergence of high-throughput technologies allowing to measure vast amounts of clinically relevant data at increasingly high speed and low

cost has scientists facing data so high in dimensionality that finding patterns and relationships among the data points exceeds the capabilities of the human brain. Meanwhile, harvesting the wealth of information contained in this data is presumed to offer new opportunities in preventing, diagnosing, and treating diseases and therefore improving the efficiency of health care and promoting the concept of precision medicine. One such data source is the human transcriptome, the whole of all transcribed RNA molecules which can be accessed by different technologies (Fig. 1a). The transcriptome has been shown to be a feature space that is well suited for Artificial Intelligence (AI)-based algorithms, and especially the blood transcriptome, which has the advantage of being relatively easily accessible, turned out to be a good predictor not only for blood-based diseases such as leukemia [1, 2], but also of other diseases, infectious [3] and noninfectious, such as cancer, [4] or neuroimmunological diseases [5, 6]. AI algorithms have the potential to detect and extract the essential information from the transcriptomic feature space, a task that is too complex to be executed by a human given that it relies on high-dimensional datasets, but also too exhaustive to be phrased in the form of classical algorithms with explicitly defined rules. Instead, AI models can quickly comprehend and process a great wealth of information from high-dimensional transcriptomic data, granting substantial advantages in time-sensitive procedures such as prediction of disease course and outcome [7–10], detecting disease subtypes [11], or guide adequate medication of a patient [10, 12, 13]. Furthermore, AI algorithms have the potential to evolve based on the data they are presented and can, in

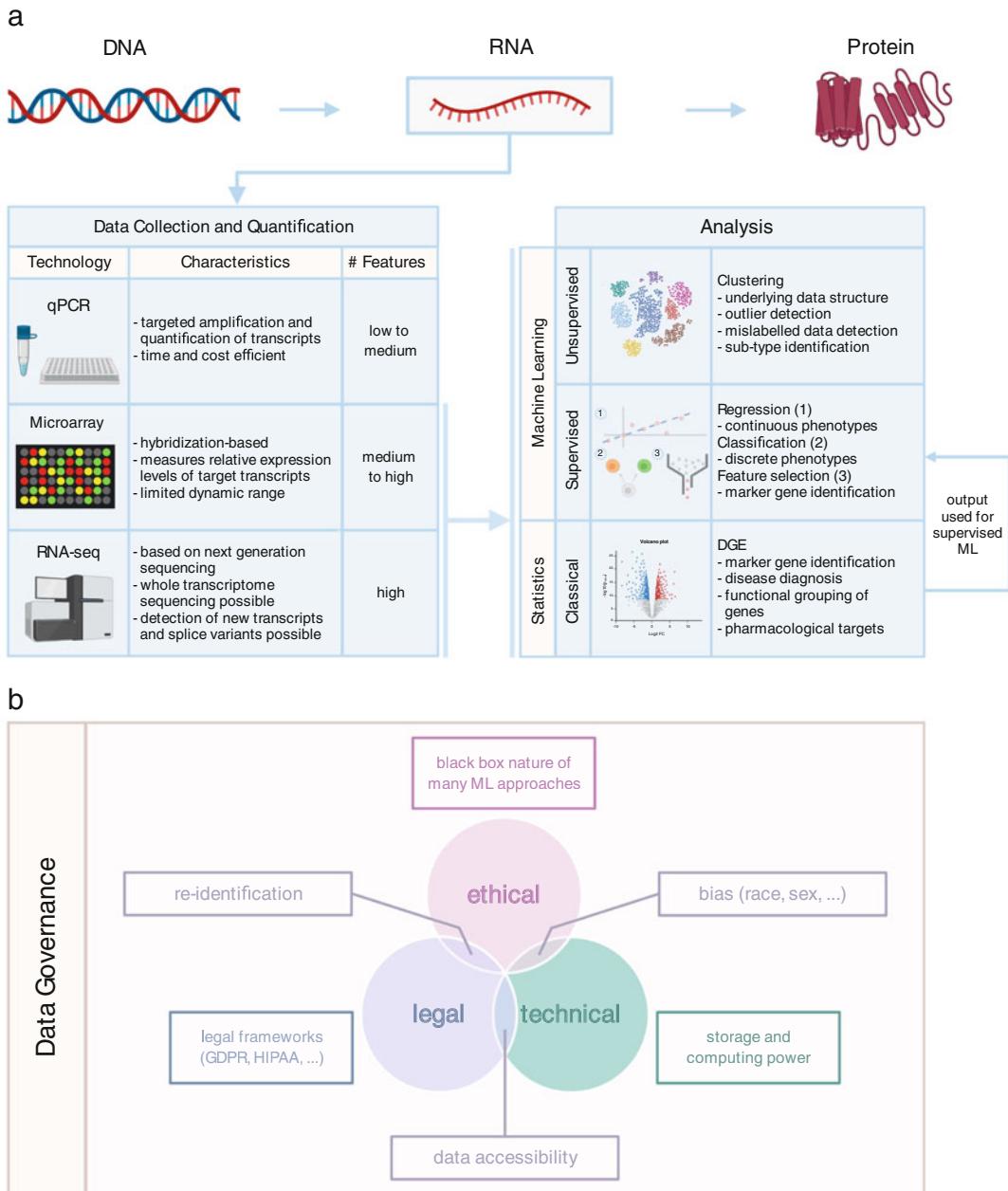


Fig. 1 Collection, analysis, and aspects for data governance of transcriptomic data. (a) Transcriptomic data represents the abundance of RNA molecules of either a selected panel of transcripts or of the whole transcriptome. The measured RNA molecules typically undergo a series of preprocessing steps (not shown) that are specific to the used protocol. Quantification is commonly conducted using qPCR, microarrays or RNA-seq methods. The

collected data is analyzed using a variety of techniques from classical statistics to machine learning. (b) Shown is a number of common issues that arise from storing and sharing transcriptomics data. Technical difficulties are mainly caused by the size of the data and the computing power required by the analytical methods. Ethical and legal issues arise due to the personal, and therefore sensitive, nature of the data

principle, continuously update and improve their predictions [13]. The speed of AI algorithms, which is unmatched by a human, could lower the costs of health care and therefore increase the amount of financial resources and medical staff available for patients in need of intensive care [13].

However, the full potential of AI-based transcriptomics to truly enhance clinical diagnosis as well as disease monitoring on a large scale is not yet translated into clinical practice and molecular diagnostics. Reasons for this include technical and infrastructure requirements that must be met at hospitals to build, train, and run models, as well as complex ethical and legal considerations (Fig. 1b).

In this chapter, we will discuss current technologies and their potential for clinical application (section “[Blood Transcriptomics in Clinics: Methods, Features, Pitfalls](#)”), different algorithms suitable for high-dimensional transcriptomic data (section “[Background on Artificial Intelligence in Biology](#)”), research development towards clinical applications (section “[Research Development Towards Clinical Applications](#)”), and important considerations in terms of data security, potential biases, and data management (section “[Ethical Considerations, Data Security and Federated Learning](#)”).

Blood Transcriptomics in Clinics: Methods, Features, Pitfalls

When translating research on transcriptomic data into AI-based clinical applications, several technical and analytical aspects need to be considered. First, the measurable feature space varies across technologies and the right method to choose is dependent on the analytical question that is being addressed. Second, downstream processing, e.g., of high-throughput RNA-sequencing data is not trivial, and standardized clinical usage of that data requires highly streamlined and reproducible pre-processing workflows which prepare the data prior to the actual prediction by AI-based methods.

Regarding the RNA features which are measured, a fundamental difference between technologies is that between a targeted approach, where a defined set of features is measured, and that of “whole transcriptome” approaches, which do not

require a priori knowledge of the features of interest (Fig. 1a).

A large body of transcriptome studies which contributed to a better understanding of disease biology are based on microarray data, which is a targeted approach to transcriptomics. Here, all RNA molecules are being converted to cDNA and labeled with fluorochrome dyes. DNA sequences complementary to the target cDNA molecules are bound to a microarray chip, to which the cDNA hybridizes. The fluorescence of the hybridized target transcripts on the chip is then measured and used as a surrogate of the transcript abundance of each target transcript. Initially, quality concerns were raised on microarray data concerning reliability and consistency [14], which led to the establishment of microarray standards, quality measures [15], and a consensus of data analysis for the development and validation of predictive models [16]. Consequently, the usage of microarray data in clinical application has been proposed for disease prediction [17–19], subtype discovery [20, 21], and subtype prediction [22]. However, the use of microarrays for accessing the transcriptome did not translate into clinics yet at a larger scale, whereas genetic microarrays to screen for chromosomal aberrations [23] and single-nucleotide polymorphisms (SNPs) [24] in patients with unspecified genetic disorders and primary immunodeficiency have become a cost-efficient standard tool.

Following more recent technological developments, microarray-based transcriptome technologies have largely been replaced by next-generation RNA sequencing (RNA-seq) protocols, which have the great advantage that they not only quantify known transcripts, but can also detect unannotated transcripts, identify alternatively spliced genes or detect allele-specific expression [25]. Compared to microarrays, RNA-seq also has a wider dynamic range of detecting transcripts, which makes it the method of choice for recent clinical transcriptomic studies. Also, with the introduction of single-cell sequencing techniques [26] it has become possible to study gene expression of individual cells, which is particularly interesting for understanding disease biology at a cellular level [27] and developing disease- and cell-type-specific predictive signatures [28].

RNA-seq protocols are manifold and can differ in the RNA species they detect. While most protocols aim at detecting polyadenylated messenger RNA (mRNA), it is also possible to investigate nonpolyadenylated RNA species, such as microRNAs, certain long noncoding RNAs, or pre-mRNA. In a nutshell, RNA is extracted and specific RNA species are being selected mostly by either enrichment of polyA-containing transcripts or the depletion of highly abundant ribosomal RNA [29]. Then, the selected RNA is fragmented, converted into cDNA, sequencing adapters are added, and the cDNA is amplified in order to construct a library for sequencing. In the case of Illumina sequencing, libraries are then loaded on a flow cell, where the cDNA molecules bind to oligonucleotides that are complementary to the sequencing adapter. Sequencing by synthesis [30] is performed with the primary sequencing output being per-cycle binary base call (BCL) files which store base calls and quality for each sequencing cycle. These are then translated into per-read FASTQ files that contain the base sequences of the library fragments, which correspond to the transcribed RNA molecules of the original specimen as well as the corresponding quality scores. In order to retain gene expression information from those fragments of sequences, they need to be mapped to a reference genome or transcriptome, depending on the application (for a detailed review see [31]). Then the relative expression level of each target feature (transcript or gene) is quantified, which results in a matrix of expression counts.

A promising alternative to next-generation sequencing protocols that use sequencing by synthesis are long-read sequencing protocols [32–34] that produce reads, e.g., in the range of 5 to 15 kb for SMRT sequencing or even ultralong (100 kb or more) nanopore reads [35]. In contrast, the maximum read length of Illumina NovaSeq 6000 is paired-end 250 bps [36]. Long read sequencing is of particular interest for applications that rely on isoform identification, for targeted sequencing of complex genomic and paralogous regions and also in particular for de novo assembly approaches [32]. This clearly illustrates that the technologies used to determine transcriptomes will have an enormous impact on further

downstream data preprocessing, data analysis, but also interpretation of the data and utility for machine learning.

Depending on the RNA-seq technology used, the amount and type of studied RNA features and based on the features that are considered during preprocessing, such count matrices can contain up to 200,000 features, which is roughly the number of annotated human transcripts, and even more if RNA-seq data is used for de novo assembly of unannotated transcripts [37]. Most current routine RNA-seq pipelines, however, aim at reducing background noise of low-expressed features before the actual biological analysis and, e.g., around 20,000 informative features are usually used for downstream analyses on the gene level of bulk RNA-seq data. Also, with decreasing costs of RNA-seq methods [38], whole-transcriptome protocols can be envisioned to become a standard tool in clinical diagnostics.

In contrast to RNA-seq and microarray application, which can both be used to generate high-dimensional datasets, quantitative reverse transcriptase PCR (qRT-PCR), a protocol which quantifies the expression of only a few target genes by amplification via the polymerase chain reaction (PCR), is considered the gold standard for the qualitative and quantitative detection of cellular mRNA, e.g., in the diagnosis of acute myeloid leukemia (AML) [39], miRNAs [40], and also genomic RNA from viruses [41, 42]. Also, RT-PCR kits are available for testing gastrointestinal, respiratory, or sexually transmitted infections. However, despite their great success in clinical applications, it has to be stated that the application of such highly sensitive but targeted protocols is by definition unsuited to detect unexpected results, with a recent example being the detection of mutations of SARS-CoV-2, which could not be detected without molecular surveillance of the viral genome [43].

Background on Artificial Intelligence in Biology

Given the magnitude and diversity of the research field, the definition of AI is oftentimes broad and unspecific. It is mostly described as a machine that

demonstrates behavior, which we would classify as intelligent, such as problem solving [44] or pattern recognition [13, 45]. A more precise distinction is that into strong and weak, also referred to as narrow, AI. Strong AI references a so-far-theoretical concept of an artificial intelligence that exhibits self-awareness and functions on the same cognitive level as the human mind. A strong AI would be able to learn autonomously and solve problems of previously unseen categories without human interference. Weak AI on the other hand is AI that specializes in mastering a specific type of task and requires a large amount of human input in the form of hyperparameter tuning and training set selection [46]. All currently existing machine learning models are of the weak type and are mostly adapted to perform tasks such as clustering of unlabelled data or pattern recognition for the use of classification and regression. The following section will specifically focus on weak AI in the context of transcriptomics.

AI Methods for Transcriptomics Analyses

Differential gene expression (DGE) is a common element in the analysis of transcriptomic data. It aims to identify groups of significantly up- or downregulated genes using statistical testing, which in the case of RNA-seq data is commonly based on a negative binomial distribution [47]. An issue that arises from the typical dimensions of RNA-seq experiments, comprising few samples but thousands of genes, is the multiple testing problem [48]. To identify differentially expressed genes, a test is conducted for each gene in the dataset, leading to a high number of expected false positives. To account for this typical issue arising from *small-n-large-p* problems, the p-values require adjustment. A common method here is the Benjamini-Hochberg procedure. The results of DGE analyses are often used as input to further statistical methods, such as testing for over- or underrepresentation of functional groups and the enrichment of gene sets or of transcription factor binding motifs. The insights are used for a vast majority of application purposes such as marker gene identification, disease diagnosis, and detection of functional gene groups or potential pharmacological targets. This classical

statistical analysis approach comes at the disadvantage of explicitly having to make assumptions with regard to the underlying data such that the correct model can be chosen. Making these assumptions is often erroneous or too simplistic and the knowledge base for the decision potentially incomplete. Among other reasons, particularly this difficulty to build an explicit model prompts the use of AI. The learning process that is involved in the construction of most AI tools is described as machine learning (ML), a model design concept that does not require explicit programming of decision rules but instead develops and improves these autonomously based on data it is provided [44]. The difference between machine learning and classical statistical approaches roots in the questions that are being asked when applying either one: While traditional statistical methods aim to find a model that explicitly describes the relationship between a set of variables, machine learning focuses on generating a model that allows for the accurate outcome prediction of previously unseen data, oftentimes without providing detailed information on the underlying model architecture itself [49].

However, classical statistical methods are by no means obsolete. They often serve as a pre-selective step to provide first subsets of genes used, for instance, in ML-based classification tasks.

The subsequently presented methods are machine learning techniques that have been recently applied in transcriptomics. These techniques can be broadly divided into supervised and unsupervised ML methods.

Supervised Learning

Supervised learning is used to build a model that receives a set of input data and based on that predicts a previously unknown output label or value. The model is initially confronted with a labelled training data set from the domain in which it is meant to predict an outcome later on. After applying the model to the data, the predicted outcome is compared to the true label of the data point. The level of discrepancy is then used to adjust the model, such that it performs better in the next iteration. This process is

repeated until convergence or until a preset prediction accuracy is reached. The final model can then be used on a data set it has not seen before to make predictions or to classify the data. If the training was successful, the model has generalized enough to accurately predict the unknown data [13].

Unsupervised Learning

Unlike supervised learning, unsupervised learning does not require training on a data set with known labels. Instead, unsupervised learning techniques attempt to group the presented, unlabeled data solely based on the data itself. The cluster labels that are assigned to the detected groups do not allow any insight into the characteristics of the data points. Unsupervised learning therefore enables the detection of patterns and outliers thereof. Since it does not rely on known labels, it allows for the detection of groups in the sample space that were not known a priori. These methods mainly differ in their definition of similarity and dissimilarity which underlie the clustering process of the sample space [50].

Feature selection

Some machine learning methods allow for the selection of a subset of features that is relevant to the task. Feature selection is especially useful in cases where not all the features are expected to be required to fulfil the machine learning task, or where the feature space is larger than the sample space, as it is the case with most high-dimensional transcriptomic datasets [51]. Here, feature selection is commonly used to address the dimensionality problems of transcriptomic data sets and to make the output human-readable and more insightful by providing a distinct set of markers that were included in the prediction or classification task [51]. Not only does this increase the amount of trust put into the model by specialists, it also allows further insight into the underlying mechanisms of the studied condition. Models that commonly do not allow the explicit extraction of features are deep learning-based models, the networks of which are too complex to pinpoint a distinct set of features used for the prediction and which are therefore often referred to as black

boxes. They are considered superior in tasks involving only very few negligible features, or where features are highly interdependent, and they therefore require a large amount of data such that the size of the sample space exceeds that of the feature space.

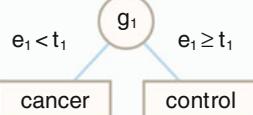
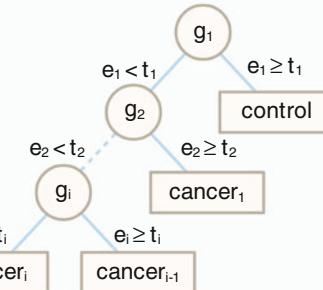
Random Forest

Random forest is a supervised machine learning technique often used for classification purposes. The name originates from the fact that the model consists of an ensemble of decision trees – a forest. A decision tree is built such that each branching point represents a decision – comprising one or several features present in the data – based on which a sample progresses along either of the two branches when traversing the tree.

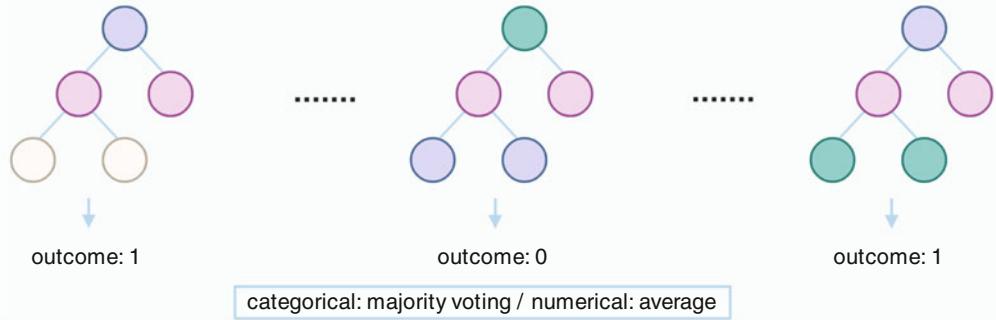
As an illustrative example consider the task of classifying a sample as belonging to either one of many specific cancer types or controls. For simplification purposes, imagine there existed a ubiquitous tumor suppressor gene g_1 where if the expression of g_1 drops below a threshold t_1 , the cell becomes tumorous without any further specification of the tumor. The expression level of g_1 could be utilized as a first branching point in a decision tree (Fig. 2a). Then, to distinguish between tumor types, imagine a distinctive marker gene g_i including a corresponding threshold t_i for each kind of tumor that would provide for further distinction along the tree (Fig. 2b).

In reality, genes that allow for such clear splits are rare and unknown, which makes the precise design of such an optimal decision tree impossible. Hence, the trees are built by randomly selecting the features for the next split, making one tree's prediction less reliable, a problem that is addressed by building an ensemble of trees. The ensemble utilizes the “wisdom of crowds” to reduce the classification error, following the idea that the majority of opinions is likely to be the correct one if the vote count is large enough. After collecting the outputs of all individual trees, the random forest classifies a given sample based on the majority of votes (Fig. 2c). In order for the ensemble to work, it is important that the single trees are as independent from another as possible. To ensure this, two techniques are typically used

Decision Trees

a**b**

Random Forest

c

Support Vector Machine

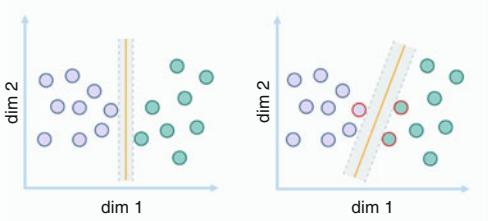
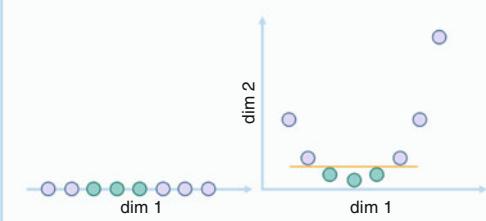
d**e**

Fig. 2 Overview of selected machine learning methods. **Decision Trees.** (a) Example for an initial branching that separates the samples into “cancer” and “control,” given the expression level e_1 of a gene g_1 . (b) After the initial categorization, cancer samples are further classified based on the expression levels of cancer-type-specific genes. **Random Forest.** (c) Node colors represent features and each tree represents a different decision sequence. The ensemble of trees predicts the final outcome either by majority voting, in the case of categorical variables, or by

averaging, in the case of numerical variables. **Support Vector Machine.** (d) The hyperplane is found by maximizing the margin. In a two-dimensional plot, the hyperplane is represented by a line. Decreasing the slope in this case increases the margin from the left to the right. Final support vectors are indicated by a red border. (e) The data points cannot be optimally separated in the one-dimensional space (left). A polynomial kernel of degree 2 transforms the data into a higher dimensional space, where the linear separation becomes possible (right)

when building a random forest: Bagging (Bootstrap Aggregating) and feature randomness.

Bagging describes the process of creating a training set for each tree by sampling from the sample space while allowing elements to be chosen more than once (sampling with replacement). This results in a data set that only contains samples that were present in the original data set, but at different frequencies. The purpose of bagging is to expose the ensemble to versions of the data that incorporate small changes, making the model more robust towards aberrations in the input data.

Feature randomness refers to the construction of the decision trees. To avoid trees from looking too similar and therefore being strongly correlated, only a subset of random features is made available for each branching point, with the consequence that the trees represent different decision sequences.

Bagging and feature randomness attempt to make the trees as uncorrelated as possible and therefore reduce the error of the model [52].

Support Vector Machine

Support Vector Machines (SVMs) are a supervised machine learning method for classifying samples into two separate classes. In order to do so, the algorithm determines a hyperplane that divides the data points into the two classes such that the margin of that hyperplane is maximal. For data points with n-features which are therefore projected into an n-dimensional space, a hyperplane defines an $(n - 1)$ -dimensional subspace [53]. The margin of that hyperplane describes the minimum distance of a data point to the hyperplane. The margin is maximized if the hyperplane is in equal distance to the closest data point from each group (Fig. 2d). Once the hyperplane is found, it represents a decision boundary for new samples: The samples will be classified based on which side of the hyperplane they are located.

If the data points are arranged in a way such that no optimally separating hyperplane can be found in their original dimension, SVMs utilize kernel functions to project the data into a higher-dimensional space in which such a separation becomes possible (Fig. 2e). A kernel function

has the advantageous attribute of providing the dot product of two vectors – which is needed to determine the hyperplane – in higher dimensions, without explicitly having to map all of the data points to that dimension [54], which would come at great computational cost [55].

In the context of transcriptomics, SVMs can be utilized to classify a new sample as one or another biological group, e.g., “case” or “control.” The size of the feature space is defined by the genes considered for building the model. This can be either the unfiltered vector of genes that were quantified during the sequencing process or a subset acquired by a feature selection method. If n is the length of every sample’s expression vector, then the SVM would attempt to find an $(n - 1)$ -dimensional hyperplane that best segregates the samples into the two labelled groups within the N -dimensional feature space.

Support Vector Machines come with one distinct disadvantage, that is their restricted ability to only solve binary classification problems.

Lasso

The Least Absolute Shrinkage and Selection Operator (LASSO) is a variant of linear regression that addresses and solves the problem of overfitting inherent to the classical linear regression of ordinary least squares by adding a regularization term to the objective function. This regularization term is an L_1 -norm, the introduction of which imposes a minimization of the least absolute errors. The consequence of the regularization term in LASSO is that the coefficients of features with low predictive weight are shrunk to 0, which introduces sparsity into the feature space and therefore is accompanied by an explicit selection of features for the model [56]. In transcriptomics analyses of whole genome sequencing data, the analyst often faces expression values for several thousands of genes, all while a standard dataset only comprises a few hundred samples – and that would already be considered rather large. Here, LASSO is a helpful tool to extract a manageable number of genes to build a more comprehensive model. Consider a case in which the gene expression profiles of different cancer types are to be

analyzed with respect to finding a model that allows the prediction of the cancer type present in a sample. It is unlikely that all genes are involved in the cancer phenotype, but rather that there exists a subset of biomarkers that allow the discrimination between the different kinds and which are therefore represented by LASSO as features with nonzero weights [1].

K-Nearest Neighbors (KNN)

The k -nearest neighbors algorithm is a supervised machine learning algorithm that can be used for both classification and regression. The underlying assumption of the algorithm is that in a given feature space, those samples that have similar features are located in closer proximity to each other than samples with vastly different features. Given a labelled set of samples and a new, unlabelled sample the label of which is to be predicted, the distance of that unlabelled sample to every labelled sample is calculated based on a predefined distance measure (e.g., Euclidean distance, Manhattan distance, Minkowski distance). Subsequently, the k labelled samples with the smallest distance are selected. In the case of a classification task, the sample is assigned a label based on the label majority among its k -nearest neighbors, the output is hence discrete. In the case of regression, the sample is assigned the mean value of its k -nearest neighbors, the output is therefore real. Given that for every sample which is to be classified, all pairwise distances to the available labelled data points must be calculated, this algorithm quickly becomes infeasible with growing sample and feature space [57]. The growth in the feature space poses a particular issue when working with unfiltered transcriptomics data, since the number of dimensions of the feature space in which the samples are located equals the number of observed genes, therefore equating to a few thousands. Thus, to reduce the size of the feature space, a preselection of genes (feature selection) or their transformation into another, summarizing feature space (feature extraction) may be necessary depending on the available computing power.

Deep Neural Networks (DNNs)

Many supervised learning approaches use artificial neural networks (ANNs) or deep ANNs (DNNs). These comprise an input layer, which represents the data fed into the model, followed by one – or more in the case of DNNs – hidden layer and eventually an output layer representing the model's predictions [13]. Each layer consists of one or more neurons, the neurons of different layers are connected by weights. In the process of forward propagation, the input is transported from the input layer through the hidden layers and finally to the output layer. After comparing the predicted outcome to the true label, a reverse signal travels from the output layer back to the input layer, adjusting weights accordingly and improving the predictive power of the network. This process is called backpropagation. If the network only propagated the outputs from one layer to the other following the scheme above, the entire network would be linear and could therefore be condensed to a two-layer network – input and output layer only – that retrieves exactly the same results. Such a network would per definition not be a *deep* neural network. What actually characterizes a DNN and differentiates it from a regular ANN is nonlinearity. Nonlinearity is introduced by applying a nonlinear function to the output of each layer before forwarding that output to the next layer. The most common nonlinear function is ReLU (Rectified Linear Unit), which sets all negative outputs to zero [58, 59].

The application of DNNs to transcriptomics data has been the subject of research in recent years and is therefore a rather young topic. Different scenarios in which DNNs could provide insight into transcriptomics data are thinkable. In a straightforward approach, the input to the DNN could be directly provided as the gene expression vector of a given sample but also other representations that require a previous transformation, for instance, into an image format [60], have been discussed. For the output layer one could imagine a phenotype classification, functional assignment [60] or – in the case of bulk RNA-seq inputs – a deconvolution of cell types [61].

Research Development Towards Clinical Applications

While approaches to ML-aided everyday diagnostics based on unstructured data have proven rather difficult – exemplified by IBM's Watson Health with the exception of Watson for Genomics [62] – approaches based on structured data such as genomics or transcriptomics data have demonstrated substantial progress in the field. Especially research into disease classification and biomarker identification has profited from ML in recent years.

The investigation of machine learning methods for disease subtype discovery and disease classification is prompted by the heterogeneity in the way in which many diseases manifest in patients, which is also mirrored in individual transcriptomic and genetic profiles. This entails that not all individuals that suffer from the same disease respond adequately to a given treatment. Machine learning techniques have been applied in order to identify subgroups of diseases in different patients to allow a more tailored treatment strategy and to gather insight into the molecular differences of disease specifications.

Already in the very first study that exemplified how high-dimensional gene expression data could be used with machine learning algorithms [2], leukemia has been a prime use case for both class discovery and class prediction and many subsequent studies have dealt with transcriptomic data from hematologic malignancies. Studies using unsupervised machine learning algorithms have identified clinically relevant subtypes and signatures [18–20, 22, 63] that are predictive of disease outcome and severity. The WHO classification of hematologic malignancies distinguishes multiple subtypes of AML that take into account cytogenetic and molecular characteristics [64], however, complete AML diagnosis and risk assessment remains limited within current diagnostic pipelines. A recent study by Arindrarto et al. [65] proposes a comprehensive whole transcriptome sequencing-based pipeline which could be used as a comprehensive platform for

differential AML diagnosis. Furthermore, primary diagnosis of acute myeloid leukemia has been shown to be feasible in a near-automated and low-cost machine learning setting based on transcriptomic data which is also robust concerning the used machine learning algorithm [1].

Disease subclassification has furthermore been performed in various other studies that dissected disease heterogeneity. Random Forest and Support Vector Machines were used to stratify patients suffering from systemic lupus erythematosus [12] and Ebola virus [10], Random Forest and Functional Trees enabled the classification of different states and stages of Multiple Sclerosis [7].

Another research focus that often goes hand in hand with subtype detection and disease classification is the development of biomarker panels. Finding a set of biomarkers specific to a disease allows for a faster diagnosis or a timely prediction of disease progression and therefore makes it easier to establish a fitting treatment plan. SVMs and Random Forest have been used in this regard to identify a set of biomarkers indicative of risk for the development of active tuberculosis [66] as well as for the development of a biomarker panel that identifies pancreatic ductal adenocarcinoma (PDAC), which differentiates the disease from chronic pancreatitis and allows for the distinction between patients with better or worse chances of survival [9]. Besides this, boosted decision trees have been used to identify a set of circulating RNAs for the identification of pregnancies at risk of inducing preeclampsia [67], and PCA-based feature extraction has allowed for the determination of a set of genes indicative of the progression of dengue hemorrhagic fever after infection with the dengue virus [8].

Ethical Considerations, Data Security and Federated Learning

Artificial intelligence, and in particular machine learning, relies on large data sets for training, and gathering them is challenging from an ethical, legal, and technical perspective (Fig. 1b).

In the experiment design phase, the selection of data sources and studies can introduce biases, with racial and sex biases among the more common ones, either through over- or under-representation of populations, age-cohorts, selection of controls, and other factors [68, 69]. In particular for rare diseases with little prevalence, creating a balanced data set can be a challenge and the results often lack statistical power or significance [70].

During the recruitment phase, patient consent needs to be adapted to include ML information and questions [71]. Legal frameworks like GDPR [72] in the EU and HIPAA [73] in the USA restrict the use of patient data and require careful planning of data processing steps [74]. All unnecessary personally identifiable information needs to be stripped from the data and pseudonyms are used. Nonetheless, re-identification and linkage attacks remain possible [75, 76]. Model inversion attacks can be exploited to recover training or patient information from models [77].

Medical data is inherently distributed data and sharing data or patient information is against medical traditions. Data generation and collection happens in physicians' offices, hospitals, or in research centers that often do not have the resources for running their own ML studies and the data sets are typically too small to train reliable models. To overcome this, data can be collected in central clouds [78] and processed there, reducing the technical requirements. However, privacy and data sovereignty concerns makes this often infeasible for patient data. Federated AI approaches [79] have been introduced where the data is kept at the source and the central cloud controls the model parameters distribution and merging. This reduces privacy concerns; however, a central single point of failure still exists. Swarm learning [80] is a decentralized approach that replaces all central components. For each epoch, the swarm elects a leader that performs the parameter merging. Participation in the swarm is controlled by blockchain-based smart contracts to ensure all legal requirements.

The success of artificial intelligence in the clinics depends on the acceptance by physicians and patients. Interpreting the output of ML approaches can be difficult if they act like a black box [81]. Some approaches like KNN or

SVMs allow easy understanding of the deciding factors, but other approaches like ANNs do not offer such easy explanations. Here, explainable AI techniques are needed to expose the underlying mechanisms and to build trust in the suggestions from AI systems.

Outlook

The combination of blood transcriptomics and AI methods has great potential to become an integral part of clinical routine. Blood transcriptomics provides a rich feature space, which is particularly suited for predictions in the context of systemic inflammatory conditions, be they infectious or noninfectious, acute or chronic. If trained correctly, without bias and under strict guidelines concerning data security, AI solutions based on strong feature spaces such as blood transcriptomes could become important supportive systems for the decision-making process by physicians. However, AI algorithms to be applied in such scenarios can all be categorized as weak AI, which leaves the final responsibility for treatment decisions and the patient's well-being with the physician. These AI solutions for primary and differential diagnosis of diseases are at the heart of Systems Medicine, which combines cutting-edge technologies of computer science and biotechnology. Without any doubt, the focus in the development of medical AI-systems should be put on technologies that build on and pursue valuable traditions of medicine such as learning from each other and the importance of patient privacy. In this regard, developments such as Swarm Learning that serve both collective learning and data privacy protection need to be fostered. A fruitful interaction between the medical and computer science communities will be an important prerequisite to achieve these goals.

Cross-References

- ▶ [AI and Immunoinformatics](#)
- ▶ [AIM in Genomics](#)
- ▶ [AIM in Haematology](#)
- ▶ [AIM in Oncology](#)

► Artificial Intelligence in Clinical Immunology
 ► Artificial Intelligence in Medicine and Privacy Preservation

Acknowledgments This work was supported in part by the German Research Foundation (DFG) –(INST 37/1049-1, INST 216/981-1, INST 257/605-1, INST 269/768-1 and INST 217/988-1, INST 217/577-1); Germany's Excellence Strategy (DFG) – (EXC2151 – 390873048); the HGF incubator grant sparse2big (ZT-I-0007); the HGF grant ProGeneGen (ZT-1-PF-5-23); the EU project SYSCID (grant number 733100); the BMBF-funded excellence project Diet–Body–Brain (DietBB) (grant number 01EA1809A); and “NaFoUniMedCovid19” (FKZ: 01KX2021, project acronym “COVIM”). All figures were created with BioRender.com.

References

- Warnat-Herresthal S, Perrakis K, Taschler B, et al. Scalable prediction of acute myeloid leukemia using high-dimensional machine learning and blood transcriptomics. *iScience*. 2020;23:100780. <https://doi.org/10.1016/j.isci.2019.100780>.
- Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. 1999;286:531–7. <https://doi.org/10.1126/science.286.5439.531>.
- Thompson EG, Du Y, Malherbe ST, et al. Host blood RNA signatures predict the outcome of tuberculosis treatment. *Tuberculosis (Edinb)*. 2017;107:48–58. <https://doi.org/10.1016/j.tube.2017.08.004>.
- Best MG, Sol N, Kooi I, et al. RNA-Seq of tumor-educated platelets enables blood-based pan-Cancer, multiclass, and molecular pathway Cancer diagnostics. *Cancer Cell*. 2015;28:666–76. <https://doi.org/10.1016/j.ccr.2015.09.018>.
- Feng X, Bao R, Li L, et al. Interferon-β corrects massive gene dysregulation in multiple sclerosis: short-term and long-term effects on immune regulation and neuroprotection. *EBioMedicine*. 2019;49:269–83. <https://doi.org/10.1016/j.ebiom.2019.09.059>.
- Lee T, Lee H. Prediction of Alzheimer’s disease using blood gene expression data. *Sci Rep*. 2020;10:3485. <https://doi.org/10.1038/s41598-020-60595-1>.
- Acquaviva M, Menon R, Di Dario M, et al. Inferring multiple sclerosis stages from the blood transcriptome via machine learning. *Cell Rep Med*. 2020;1:100053. <https://doi.org/10.1016/j.xcrm.2020.100053>.
- Taguchi YH. Principal components analysis based unsupervised feature extraction applied to gene expression analysis of blood from dengue Haemorrhagic fever patients. *Sci Rep*. 2017;7:44016. <https://doi.org/10.1038/srep44016>.
- Khatri I, Bhasin MK. A transcriptomics-based meta-analysis combined with machine learning identifies a secretory biomarker panel for diagnosis of pancreatic adenocarcinoma. *Front Genet*. 2020;11:572284. <https://doi.org/10.3389/fgene.2020.572284>.
- Liu X, Speranza E, Muñoz-Fontela C, et al. Transcriptomic signatures differentiate survival from fatal outcomes in humans infected with Ebola virus. *Genome Biol*. 2017;18:4. <https://doi.org/10.1186/s13059-016-1137-3>.
- Aschenbrenner AC, Mouktaroudi M, Krämer B, et al. Disease severity-specific neutrophil signatures in blood transcriptomes stratify COVID-19 patients. *Genome Med*. 2021;13:7. <https://doi.org/10.1186/s13073-020-00823-5>.
- Figgett WA, Monaghan K, Ng M, et al. Machine learning applied to whole-blood RNA-sequencing data uncovers distinct subsets of patients with systemic lupus erythematosus. *Clin Transl Immunol*. 2019;8: e01093. <https://doi.org/10.1002/cti2.1093>.
- Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng*. 2018;2:719–31. <https://doi.org/10.1038/s41551-018-0305-z>.
- Marshall E. Getting the noise out of gene arrays. *Science*. 2004;306:630–1. <https://doi.org/10.1126/science.306.5696.630>.
- MAQC Consortium, Shi L, Reid LH, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*. 2006;24:1151–61. <https://doi.org/10.1038/nbt1239>.
- Shi L, Campbell G, Jones WD, et al. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat Biotechnol*. 2010;28:827–38. <https://doi.org/10.1038/nbt.1665>.
- van ‘t Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002;415:530–6. <https://doi.org/10.1038/415530a>.
- Kuiper R, Broyl A, de Knegt Y, et al. A gene expression signature for high-risk multiple myeloma. *Leukemia*. 2012;26:2406–13. <https://doi.org/10.1038/leu.2012.127>.
- Zhan F, Barlogie B, Arzoumanian V, et al. Gene-expression signature of benign monoclonal gammopathy evident in multiple myeloma is linked to good prognosis. *Blood*. 2007;109:1692–700. <https://doi.org/10.1182/blood-2006-07-037077>.
- Bullinger L, Döhner K, Bair E, et al. Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *N Engl J Med*. 2004;350:1605–16. <https://doi.org/10.1056/NEJMoa031046>.
- Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. 2000;403:503–11. <https://doi.org/10.1038/35000501>.
- Andersson A, Ritz C, Lindgren D, et al. Microarray-based classification of a consecutive series of 121 childhood acute leukemias: prediction of leukemic and genetic subtype as well as of minimal residual disease status. *Leukemia*. 2007;21:1198–203. <https://doi.org/10.1038/sj.leu.2404688>.

23. Miller DT, Adam MP, Aradhya S, et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am J Hum Genet.* 2010;86:749–64. <https://doi.org/10.1016/j.ajhg.2010.04.006>.
24. Surattanond N, van Wijck RTA, Broer L, et al. Rapid low-cost microarray-based genotyping for genetic screening in primary immunodeficiency. *Front Immunol.* 2020;11:614. <https://doi.org/10.3389/fimmu.2020.00614>.
25. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet.* 2009;10:57–63. <https://doi.org/10.1038/nrg2484>.
26. Potter SS. Single-cell RNA sequencing for the study of development, physiology and disease. *Nat Rev Nephrol.* 2018;14:479–92. <https://doi.org/10.1038/s41581-018-0021-7>.
27. Schulte-Schrepping J, Reusch N, Paclik D, et al. Severe COVID-19 is marked by a dysregulated myeloid cell compartment. *Cell.* 2020;182:1419–1440.e23. <https://doi.org/10.1016/j.cell.2020.08.001>.
28. Bernardes JP, Mishra N, Tran F, et al. Longitudinal multi-omics analyses identify responses of megakaryocytes, erythroid cells, and Plasmablasts as hallmarks of severe COVID-19. *Immunity.* 2020;53:1296–1314.e9. <https://doi.org/10.1016/j.jimmuni.2020.11.017>.
29. Zhao S, Zhang Y, Gamini R, et al. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Sci Rep.* 2018;8:4781. <https://doi.org/10.1038/s41598-018-23226-4>.
30. Ju J, Kim DH, Bi L, et al. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc Natl Acad Sci USA.* 2006;103:19635–40. <https://doi.org/10.1073/pnas.0609513103>.
31. Van den Berge K, Hembach KM, Soneson C, et al. RNA sequencing data: hitchhiker's guide to expression analysis. *Annu Rev Biomed Data Sci.* 2019; <https://doi.org/10.1146/annurev-biodatasci-072018-021255>.
32. Pollard MO, Gurdasani D, Mentzer AJ, et al. Long reads: their purpose and place. *Hum Mol Genet.* 2018;27:R234–41. <https://doi.org/10.1093/hmg/ddy177>.
33. Bowden R, Davies RW, Heger A, et al. Sequencing of human genomes with nanopore technology. *Nat Commun.* 2019;10:1869. <https://doi.org/10.1038/s41467-019-09637-5>.
34. Ardui S, Ameur A, Vermeesch JR, Hestand MS. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res.* 2018;46:2159–68. <https://doi.org/10.1093/nar/gky066>.
35. Amarasinghe SL, Su S, Dong X, et al. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* 2020;21:30. <https://doi.org/10.1186/s13059-020-1935-5>.
36. Illumina Maximum read length for Illumina sequencing platforms. <https://support.illumina.com/bulletins/2020/04/maximum-read-length-for-illumina-sequencing-platforms.html>. Accessed 16 Feb 2021.
37. Morillon A, Gautheret D. Bridging the gap between reference and real transcriptomes. *Genome Biol.* 2019;20:112. <https://doi.org/10.1186/s13059-019-1710-7>.
38. Alpem D, Gardeux V, Russeil J, et al. BRB-seq: ultra-affordable high-throughput transcriptomics enabled by bulk RNA barcoding and sequencing. *Genome Biol.* 2019;20:71. <https://doi.org/10.1186/s13059-019-1671-x>.
39. Haferlach T, Schmidts I. The power and potential of integrated diagnostics in acute myeloid leukaemia. *Br J Haematol.* 2020;188:36–48. <https://doi.org/10.1111/bjh.16360>.
40. Forero DA, González-Giraldo Y, Castro-Vega LJ, Barreto GE. qPCR-based methods for expression analysis of miRNAs. *BioTechniques.* 2019;67:192–9. <https://doi.org/10.2144/btn-2019-0065>.
41. Corman VM, Landt O, Kaiser M, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Euro Surveill.* 2020; <https://doi.org/10.2807/1560-7917.ES.2020.25.3.2000045>.
42. Corman VM, Müller MA, Costabel U, et al. Assays for laboratory confirmation of novel human coronavirus (hCoV-EMC) infections. *Euro Surveill.* 2012; <https://doi.org/10.2807/ese.17.49.20334-en>.
43. Cyranoski D. Alarming COVID variants show vital role of genomic surveillance. *Nature.* 589:337–8. <https://doi.org/10.1038/d41586-021-00065-4>.
44. Meskó B, Görög M. A short guide for medical professionals in the era of artificial intelligence. *npj Digital Med.* 2020;3:126. <https://doi.org/10.1038/s41746-020-00333-z>.
45. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25:44–56. <https://doi.org/10.1038/s41591-018-0300-7>.
46. IBM Cloud Education (2020) Strong AI. In: Strong AI. <https://www.ibm.com/cloud/learn/strong-ai#:~:text=Weak%20AI%2C%20also%20known%20as,to%20solve%20for%20new%20problems>. Accessed 12 Feb 2021.
47. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15:550. <https://doi.org/10.1186/s13059-014-0550-8>.
48. Johnstone IM, Titterington DM. Statistical challenges of high-dimensional data. *Philos Transact A Math Phys Eng Sci.* 2009;367:4237–53. <https://doi.org/10.1098/rsta.2009.0159>.
49. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. *Nat Methods.* 2018;15:233–4. <https://doi.org/10.1038/nmeth.4642>.
50. Trask AW. Chapter 2. Fundamental concepts: how do machines learn? In: *Grokking deep learning*. Shelter Island: Manning; 2019.

51. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*. 2007;23:2507–17. <https://doi.org/10.1093/bioinformatics/btm344>.
52. Breiman L (2001) Random Forests. *Machine Learning*.
53. Albon C (2018) 17. Support Vector Machines. *Machine Learning with Python Cookbook*.
54. Strang G. VII.5: the world of machine learning. *Linear Algebra and Learning from Data*; 2019. p. 414.
55. Huang S, Cai N, Pacheco PP, et al. Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics Proteomics*. 2018;15: 41–51. <https://doi.org/10.21873/cgp.20063>.
56. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc Series B Stat Methodol*. 1996;58:267.
57. Albon C (2018) 15. K-Nearest Neighbors. *Machine Learning with Python Cookbook*.
58. Strang G. VII.1 The construction of deep neural networks. *Linear Algebra and Learning from Data*; 2019. p. 375.
59. Trask AW. Chapter 6. Building your first deep neural network: introduction to backpropagation. In: *Grokkering deep learning*. Shelter Island: Manning; 2019.
60. Yuan Y, Bar-Joseph Z. Deep learning for inferring gene relationships from single-cell expression data. *Proc Natl Acad Sci USA*. 2019; <https://doi.org/10.1073/pnas.1911536116>.
61. Menden K, Marouf M, Oller S, et al. Deep learning-based cell composition analysis from tissue expression profiles. *Sci Adv*. 2020;6:eaba2619. <https://doi.org/10.1126/sciadv.aba2619>.
62. Strickland E (2019) How IBM Watson Overpromised and Underdelivered on AI Health Care. <https://spectrum.ieee.org/biomedical/diagnostics/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care>. Accessed 12 Feb 2021.
63. Yeoh E-J, Ross ME, Shurtliff SA, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*. 2002;1:133–43. [https://doi.org/10.1016/s1535-6108\(02\)00032-6](https://doi.org/10.1016/s1535-6108(02)00032-6).
64. Arber DA, Orazi A, Hasserjian R, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood*. 2016;127:2391–405. <https://doi.org/10.1182/blood-2016-03-643544>.
65. Arindarto W, Borras DM, de Groen RAL, et al. Comprehensive diagnostics of acute myeloid leukemia by whole transcriptome RNA sequencing. *Leukemia*. 35: 47–61. <https://doi.org/10.1038/s41375-020-0762-8>.
66. Zak DE, Penn-Nicholson A, Scriba TJ, et al. A blood RNA signature for tuberculosis disease risk: a prospective cohort study. *Lancet*. 2016;387:2312–22. [https://doi.org/10.1016/S0140-6736\(15\)01316-1](https://doi.org/10.1016/S0140-6736(15)01316-1).
67. Munchel S, Rohrback S, Randise-Hinchliff C, et al. Circulating transcripts in maternal blood reflect a molecular signature of early-onset preeclampsia. *Sci Transl Med*. 2020; <https://doi.org/10.1126/scitranslmed.aaz0131>.
68. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med*. 2018;178:1544–7. <https://doi.org/10.1001/jamainternmed.2018.3763>.
69. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA*. 2019; <https://doi.org/10.1001/jama.2019.18058>.
70. Mitani AA, Haneuse S. Small data challenges of studying rare diseases. *JAMA Netw Open*. 2020;3: e201965. <https://doi.org/10.1001/jamanetworkopen.2020.1965>.
71. Cohen IG. Informed consent and medical artificial intelligence: what to tell the patient? *SSRN J*. 2020; <https://doi.org/10.2139/ssrn.3529576>.
72. McCall B. What does the GDPR mean for the medical community? *Lancet*. 2018;391:1249–50. [https://doi.org/10.1016/S0140-6736\(18\)30739-6](https://doi.org/10.1016/S0140-6736(18)30739-6).
73. Kels CG. HIPAA in the era of data sharing. *JAMA*. 2020;323:476–7. <https://doi.org/10.1001/jama.2019.19645>.
74. Shabani M, Marelli L. Re-identifiability of genomic data and the GDPR: assessing the re-identifiability of genomic data in light of the EU general data protection regulation. *EMBO Rep*. 2019; <https://doi.org/10.15252/embr.201948316>.
75. Backes M, Berrang P, Bieg M, et al. Identifying personal DNA methylation profiles by genotype inference. 2017 IEEE symposium on Security and Privacy (SP). IEEE; 2017. p. 957–76.
76. Raisaro JL, Tramèr F, Ji Z, et al. Addressing Beacon re-identification attacks: quantification and mitigation of privacy risks. *J Am Med Inform Assoc*. 2017;24: 799–805. <https://doi.org/10.1093/jamia/ocw167>.
77. Fredrikson M, Lantz E, Jha S, et al. Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing. *Proc USENIX Secur Symp*. 2014;2014:17–32.
78. Ping P, Hermjakob H, Polson JS, et al. Biomedical informatics on the cloud: a treasure hunt for advancing cardiovascular medicine. *Circ Res*. 2018;122:1290–301. <https://doi.org/10.1161/CIRCRESAHA.117.310967>.
79. Kaassis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat Mach Intell*. 2020; <https://doi.org/10.1038/s42256-020-0186-1>.
80. Warnat-Herresthal S, Schultze H, Shastry KL, et al. Swarm Learning for decentralized and confidential clinical machine learning. *Nature*. 2021;594:265–270. <https://doi.org/10.1038/s41586-021-03583-3>.
81. Azodi CB, Tang J, Shiu S-H. Opening the black box: interpretable machine learning for geneticists. *Trends Genet*. 2020;36:442–55. <https://doi.org/10.1016/j.tig.2020.03.005>.



AIM in Health Blogs

81

Paola Velardi and Andrea Lenzi

Contents

Introduction	1126
Related Studies	1127
Social Analytics for Healthcare	1127
Adverse Drug Reaction (ADR)	1127
Methodology	1128
Data Gathering	1128
Data Filtering	1129
Topic Modeling	1130
Evaluation and Experiments	1132
Dataset and Diseases	1132
Compared Systems	1132
Evaluation Strategy	1133
Topic Analysis and Subclustering	1134
Data Analytics	1136
Conclusion	1140
References	1140

Abstract

People are generating an enormous amount of social data to describe their health care experiences, and continuously search information about diseases, symptoms, diagnoses, doctors, treatment options, and medicines. This data can be used to extract the “social phenotype”

of diseases, i.e., a *subjective and unfiltered picture of perceptions, health conditions, feelings, and lifestyles* of people affected by a given disease, taking into account the experience lived outside the traditional syndromic care and monitoring structures, the symptoms that are perceived as more limiting, the greater or lesser tolerance to drugs. In this chapter we present an innovative approach based on an Auto-ML pipeline to optimize the classification of relevant messages and threads, and on a cooperative deep learning topic detection model, to characterize the social phenotype of

P. Velardi (✉) · A. Lenzi
Department of Computer Science, Sapienza University of Rome, Rome, Italy
e-mail: velardi@di.uniroma1.it; lenzi@di.uniroma1.it

patients affected by widespread diseases, such as diabetes and hypothyroidism.

Keywords

Social phenotype of diseases · Health blogs · Quality of life assessment · Adverse drug reaction · Topic modeling · Topic clustering · AutoML · Cooperative Neural Networks

Introduction

Despite the progress made in understanding a wide range of diseases, the effectiveness of a treatment depends to a great extent on the communication between doctor and patient [1, 2]. *Patient centeredness* in doctor-patient communication has been advocated in many studies (for example, [1, 2]). Patient centeredness has been defined in [3] as “a move away from a disease focus to personalizing care according to patients concerns and preferences, considering the biological, psychological, and social aspects of illness.”

In medical literature, two methods are commonly used to improve the understanding of problems encountered by patients during a treatment: Quality of Life (QoL) assessments, and specific training programs, called “repairs.”

- Quality of Life assessments [1, 4] are survey studies in which questions are posed to patients using interviews and/or questionnaires.
- Repairs [3, 5] are training programs to improve clinicians’ communication mechanisms and develop a shared understanding of a disease experience.

Both these methods, despite their undisputed effectiveness, have also significant drawbacks. The first problem which is common to both approaches is coverage: questionnaires and training programs can be delivered to a limited number of individuals. Furthermore, repair programs are limited by lack of time and resources, while questions posed in QoL assessments often reflect the care-givers’ view of factors that mostly influence the health status, rather than the patient perspective.

In this chapter we propose to complement the aforementioned approaches with a text-mining methodology to extract from health-related forums and social networks a patient-centered perspective of diseases and treatments. Online patient health data is increasing exponentially. Every day, through social networks, search engines, online forums, web sites, mobile applications, IoT devices, etc., people generate an enormous amount of health information [6]. This digital data includes symptoms, medical experiences, bio data, biometric information, web search queries, personal stories and medical records. They represent the footprints of the patients health status and foster large-scale and low-cost analysis of common diseases [7] for many applications, such as syndromic surveillance [8] and adverse drug reactions [9]. Leveraging social data, it is possible to obtain a subjective and unfiltered picture of perceptions, health conditions, feelings and lifestyles of people affected by a given disease, that we name the *social phenotype* of a disease.

Whether the objective is to assess the quality of life of patients affected by a given disease, or their reactions to drugs, characterizing the social phenotype implies the design of accurate tools to extract and analyze the main topics of discussion in health blogs. Towards this aim, we present a framework based on an Auto-ML pipeline to identify and classify social messages and an innovative technique (*Generative Text Compression with Agglomerative Clustering Summarization*, GTCACS) to identify the main topics of discussions. The distinctive feature of our approach as compared with other topic detection models, is that we are able to organize topics in three dimensions: comorbidities and conditions, emotional states, and symptoms. Clustering patients’ messages along these dimensions may provide useful and large scale information to support doctors in quality of life assessments and therapy analysis.

The chapter is organized as follows: In section “[Related Studies](#)” we summarize related literature. Section “[Methodology](#)” presents our framework to the social phenotype of diseases. Next, we describe all the steps of the proposed workflow in more detail. In section “[Data Gathering](#)” we describe the process of data collection from the

selected web forum. Section “[Data Filtering](#)” presents the techniques implemented to filter significant messages from the collection of patients’ threads, and section “[Topic Modeling](#)” provides details on the approach used for topic modeling. In sections “[Evaluation and Experiments](#)” and “[Data Analytics](#)” we perform a quantitative and qualitative evaluation of the proposed methodology, applied to the case of diabetic and hypothyroid patients. Finally, section “[Conclusion](#)” presents a summary of findings and concluding remarks.

Related Studies

Our work is related to two main areas: social analytics for healthcare, and adverse drug reaction.

Social Analytics for Healthcare

More and more often patients search for information about diseases, symptoms, diagnoses, doctors, treatment options and medicines [6, 7, 10, 11]. A recent study states that in 2017, out of the total American social network users, 26% have discussed health information, and, of those, 30% changed behavior based on this information and 42% discussed their current medical conditions (<http://www.pewinternet.org/fact-sheet/social-media/>). According to [11], users who go online to connect with others with similar health concerns, are especially those who have chronic conditions such as high blood pressure, diabetes, and cancer. Frequently, they also search for support from online communities [12]. The increasing availability of this information presents an interesting opportunity to enhance timeliness and efficiency of care. Greaves et al. [7] describe this growing body of information on the Internet as a “cloud of patient experience.” Several studies have shown that it is possible to detect patterns in data gathered from health forums or social networks, in order to support medical decisions or predictions. There is an ample literature that demonstrates the possibility to analyze different types of discussion forums or

blogs for obtaining useful information. For example, [13] present a survey of qualitative research studies. In [14] the authors published studies on content analysis of online cancer communities, considering both qualitative and quantitative approaches. Quantitative methods, based on text mining and machine learning methods (see [15] for a survey), have been mostly concerned with the use of social networks (especially Twitter) for syndromic surveillance, pharmacovigilance and behavioral medicine, like smoke and diet behaviors (among the others, [8, 9]). In these literature reviews on patients’ blogs analytics, the majority of studies concentrate on classification problems. Fewer articles analyze more in detail users’ messages, for example, [16, 17]. In [17] the authors examined approximately 1000 posts from a public Facebook group on multiple sclerosis, manually labeling the main discussion topics and then using the extracted topic-related terms to extend their analysis to a larger sample. Information and awareness was the most discussed topic, followed by event advertising and petitions, fundraising, and patient support. Finally, in [16], the authors analyzed a database of feedback from patients who received healthcare in hospitals across United Kingdom.

Adverse Drug Reaction (ADR)

Pharmacosurveillance is mostly conducted by clinicians, however, as also reported in a recent survey on ADR [18], patients have the tendency to downgrade the severity of their symptoms, especially when involving depression and other psychological conditions. Social media offer the opportunity to obtain a more detailed and “nuanced” vision of patients and how they react to a therapy, the experience lived outside the traditional syndromic care and monitoring structures, the symptoms that are perceived as more limiting and the greater or lesser tolerance to drugs. A number of methods have been proposed to extract, analyze and classify ADR in social media [19–21]. We refer to [18] for an accurate survey. More recently, many ADR classifiers have leveraged deep machine learning methods in combination with NLP techniques. For example, in [22] the authors propose a LSTM architecture

augmented with a transfer learning approach, in which they make use of a pre-trained sentiment model to improve the quality of ADR classification on social media posts. Word embeddings have been widely used to cope with language inconsistencies, thus improving the rate of detected ADR messages (see [23, 24], among the others). Convolutional Recurrent Neural Networks have been shown in [25] to outperform other methods in the task of classifying ADR sentences in Twitter and in a dataset of MEDLINE case reports.

We note that in both application areas, available studies do not address an in-depth analytics of patients' status, as we do. The social phenotype of a disease does not simply consist in classifying positive or negative feelings about the experience of a disease, or detecting the presence of adverse reactions to drugs. Rather, it implies analyzing the main discussion topics around a disease or therapy, use this information for patient stratification into sub-cohorts, and, for each cohort, evaluate patients' nuanced feelings and reactions.

Methodology

Figure 1 summarizes the workflow to extract, filter and process online patients' messages for extracting the social phenotype. Each step is described in detail in sections “Data Gathering,” “Data Filtering,” “Topic Modeling,” and “Data Analytics.”

1. First of all, appropriate sources of data related to a disease d of interest are identified;
2. Subsequently, useful data are gathered from the selected sources and indexed, to build our collection (section “Data Gathering”);

3. Next, patients' messages related with d are extracted from the collection, using an AutoML classification pipeline to fine-tune the process (section “Data Filtering”);
4. Afterwards, patients' messages are analyzed to detect relevant discussion topics (section “Topic Modeling”), using an innovative technique named Generative Text Compression with Agglomerative Clustering Summarization. Detected topics allow to characterize patients' experience according to different dimensions, such as comorbidities, conditions, symptoms, emotional states;
5. Finally (section “Data Analytics”), topics are analyzed along the detected dimensions and according to specific application objectives, to characterize the social phenotype of d . In this chapter we shortly discuss two applications: quality of life assessment of diabetic patients, and therapy analytics for hypothyroid patients.

Data Gathering

As summarized in section “Related Studies” several studies have collected data on ADR and other biomedical applications using popular social media such as *Twitter*, *Facebook*, *Instagram* and *Reddit*. In these social media, health conditions are reported by the users in a single short message using simpler, naive terms [26], since the intent is to share their worries and feelings with other peers (e.g., “*i feel so lethargic hhh maybe if i like ate better and went and made an appt with a pcp and went back on synthroid i would feel idk better*”). We believe that, in order to gain useful insight into a patient everyday experience of a disease, health forums are a better source of information, since

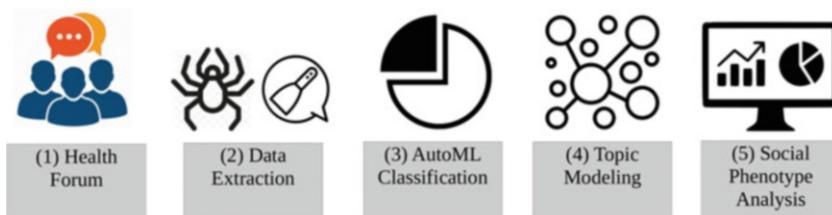


Fig. 1 Process workflow: data sources identification, data gathering, data classification, topics extraction, analysis of the social phenotype of diseases

patients provide more details on their conditions and on the effect of drugs. In health forums, patients engage a conversation with doctors on specific aspects of their health status, to receive advice. Contrary to physical visits, here the conversation is initiated and driven by patients, is centered around their symptoms and worries, and patients feel less intimidated than during a face-to-face medical examination.

Conversation threads are made of several messages and messages are composed by several sentences, written in a (relatively) technical language, since patients strive to use appropriate medical terms with their counterparts doctors. An example of message excerpt involving an ADR is: “*Currently I am taking a eutirox 175 tablet (...). My problem is that for about 4 years and only 6 years after the surgery, I suffer from continuous dizziness (as if I was always in a state of light drunkenness, disorientation, loss of concentration) associated or alternating with muscle pain of the legs (sometimes like the sensation of lactic acid).*”

Although the use of a quasi-technical language may simplify some of the problems often encountered with ungrammatical posts in social media, analyzing health forums is also quite challenging. First, given that threads are often made of several messages and span over different topics and possible comorbidities, identifying relevant messages is far from being trivial. Second, extracting aspect-based features to assess the emotional intensity around a disease experience is more complex in the context of a doctor-patient dialogue than in a short message, since a patient may report issues related to, for example, an adverse drug reaction, in a different phrase or message than the one in which he/she mentions the drug intake.

Data Filtering

After collecting and indexing a suitable number of messages from selected sources, the next step is to filter the information of interest from the collection. In fact, although in health forums threads are often manually classified by the forum curators in

several disease categories (e.g., *thyroid* or *diabetes*), this is not sufficient to identify relevant conversations. The problem is both precision and recall: we observed that often, threads explicitly tagged as related to a disease d were not truly concerned with d , nor with a therapy experience. Conversely, there are threads not tagged as related to d , where patients affected by other diseases mention a comorbidity with d , or mention the intake of d -related medications. Therefore, an accurate classification of health-related threads concerned with a disease d requires a more complex classification phase. To this end, we proceed as follows:

- Step 1: we generate a highly reliable training and test set from a subsample of the collection, by selecting messages with at least 3 highly related keywords identified by a specialist;
- Step 2: we train an optimized classifier using an AutoML pipeline;
- Step 3: we apply the optimized classifier to filter the entire collection, thus obtaining our final dataset of patients’ messages related to d .

In Step 1, based on clinical experience and the online context, we first identify a set of keywords K highly associated with the pathology under study. For example, if $d = \text{thyroid}$, relevant keywords are *euthyrox*, *hashimoto*, *hypothyroid*, *cytomegalovirus*, etc. Usually, about 100–200 keywords can be identified, depending upon the disease and purpose of analysis (e.g.: analyzing the quality of life of d patients, detecting common ADRs, classifying their mental states, and others). To train a classifier, we retrieve from the initial collection all the threads with at least three occurrences in K , and we mark them with the label d (positive examples of threads concerning d). Subsequently, we recover from the entire corpus the threads with not even one occurrence in K , we arbitrarily selected from these threads a sub-sample of messages, and we marked them with the label “other” (negative examples). At the end of this phase, we obtain a highly reliable, although possibly small, labeled set of messages, which constitutes the learning set $L(d)$ of the classifier. As a common practice, $L(d)$

is randomly split into two parts: training set (70%) and test set (30%) (the process can be repeated in k folds, depending upon the dimension of $L(d)$).

In Step 2, we determine the optimal pipeline for the classification task, adopting an automated Machine Learning (Auto-ML [27]) approach based on Sequential Model Based Optimization [28]. An AutoML system is a component that attempts to automate the end-to-end process of a Machine Learning task. It is designed to help machine learning engineers, by automatically searching through the complex space of the learning pipeline and hyperparameter settings, to maximize performance.

Finally, in Step 3 we create the final collection of threads for a disease d using the optimized classification model generated in Step 2. To this end, we consider again all the threads belonging to the entire unfiltered collection, including at least one occurrence of the initial highly related keywords. These messages constitute the “prediction set” P . Subsequently, we generate predictions applying the optimized AutoML classifier on P . Every message $m \in P$ is labeled by the classifier with a probability $p(m)$ of belonging to class d . Since we are more focused on obtaining a highly precise dataset of d -related conversations, rather than a high recall, we consider only those messages in P such that $p(m) \geq 0.9$. This process eventually results in a final collection $C(d)$ of d -related messages.

Topic Modeling

Topic modeling represents a popular technique to learn the thematic structure of large collections of unlabeled documents, without human supervision. It provides a convenient method for dimensionality reduction and exploratory data analysis in vast text bodies. Using contextual evidence, this method can connect words with similar meanings and discover the semantic structure of a collection by examining statistical co-occurrence patterns within a text corpus.

To characterize the therapy and everyday life experience of d patients, we extract the discussion topics that emerge from user messages associated

with the pathology d under study. Toward this aim, we designed a topic detection strategy, Generative Text Compression with Agglomerative Clustering Summarization (*GTCACS*) to identify high-quality discussion topics in the collection.

The proposed *GTCACS* approach is in three steps, detailed below. The *GTCACS* project is available under MIT license and can be downloaded from <https://pypi.org/project/gtcacs/>.

Dimensionality Reduction

Starting from the concept of Energy-based Generative Adversarial Network (EBGAN) suggested by Zhao et al. in [29] and from the architecture proposed by Glover in [30], we developed a model to extract the most significant latent features and learn compressed representations from unlabeled natural language documents in a collection. Generative adversarial networks (GANs) are generative architectures in which two adversarial neural nets, a discriminator and a generator, are trained simultaneously. The discriminator is trained to distinguish real samples of a dataset from fake samples constructed by the generator, while, at the same time, the generator is trained to produce synthetic samples from a random source, in order to train the discriminator to distinguish between the fake generated data and the real data. In the original formulation of GAN (Goodfellow et al. [31]), the discriminator was a probabilistic binary classifier with logistic output, and convergence occurred when the distribution produced by the generator matches the data distribution. More recently, Zhao et al. In [29] propose an innovative model called Energy-based Generative Adversarial Network (EBGAN) that exhibit more stable behavior during training, enabled to implement an extensive variety of Neural Networks architectures and loss functions for the discriminator, and shows higher generative capabilities than regular GANs. In the EBGANs, the generator is trained to produce contrasting samples with minimal energies, while the discriminator represents an energy function (a cost function that maps each point of an input space to a single scalar, which is called “energy”) that is trained to attribute low energies to the “real” data and higher energies to the generated synthetic samples.

Building on the previously illustrated techniques, we propose a model consisting of two neural nets which, instead of acting adversely, are trained simultaneously in a cooperative way. Our model represents a hybrid approach between an EBGAN and a denoising autoencoder and is composed by the followings two networks:

- The generative network learns to produce synthetic documents, with the aim of generating samples that are easily represented by the discriminator and tend to be similar to real data. In our context, a document is represented by a d -related patient's message. The generator has the primary objective of adaptively introducing noise into the training process of the two networks, for regularizing them, and consequently increasing the model's resistance to over-fitting.
- The non-generative network (discriminator) is an auto-encoder. In the first place, it learns how to efficiently compress the input data (both real and synthetic samples) into a lower-dimensional space. Subsequently, it learns how to reconstruct the input data back from the reduced encoded latent space, to a representation that is as close to the original input as possible. As cost function, we adopted the negative value of cosine similarity between the input data and the reconstructed output data.

Let X be a real data sample, $G(Z)$ a synthetic sample generated from a random source Z by the generator, and $D(s)$ the reconstructed sample returned by the discriminator (in our experiments, $G(Z)$ represents a single synthetic message)); then the discriminator loss L_D and the generator loss L_G are formally defined by:

$$L_G(Z) = -\cos_sim\{G(Z), D(G(Z))\} \quad (1)$$

$$\begin{aligned} L_D(G(Z), X) = & -\cos_sim\{G(Z), D(G(Z))\} \\ & - \cos_sim\{X, D(X)\} \end{aligned} \quad (2)$$

The architecture of the model is showed in Fig. 2. Generator and Discriminator were optimized using two separate instances of Adam with a learning rate of 0.005 and a β_1 value of 0.5. The β_1 value was suggested in [32], and, similarly to what reported by the authors, seems to stabilize the training phase in our model. Details on the network structure and hyper-parameters can be found in the previously indicated project link.

Agglomerative Hierarchical Clustering

Once the data samples were reduced to a compressed form containing the essential and most significant features of the text documents, we

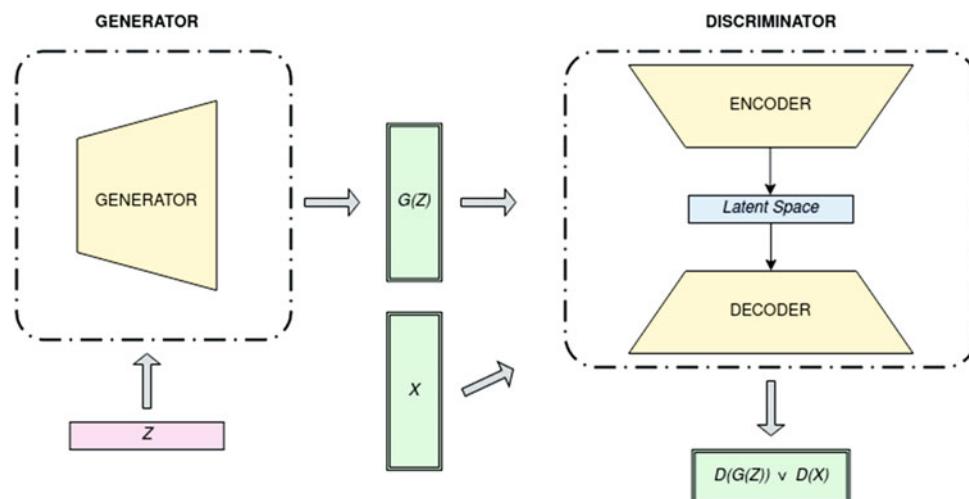


Fig. 2 Architecture of the proposed Cooperative Neural Networks for text compression

proceeded to generate document clusters, based on these representations. We performed various simulations in which we tested several off-the shelf algorithms with different number of clusters, and we evaluated the obtained performance. At the end of this experimental phase, we established to apply agglomerative hierarchical clustering with Ward method [33] as linkage criterion.

Summarization

After determining the groups of highly related documents, we extracted the most representative words for each of these clusters. Specifically, for each subset of documents associated to a single cluster, we extracted the most relevant words for this subset based on a custom weighting scheme. After various empirical tests, for each word w belonging to cluster C , we considered as the relevance score for w in C the ratio between the frequency of the term w within the documents belonging to cluster C and the global frequency of the term w in the whole corpus. In particular, we applied the following formula:

$$\text{relevance_score}(w, C) = \frac{\text{inner_term_frequency}(w, C)}{\sqrt{\text{global_term_frequency}(w)}} \quad (3)$$

Evaluation and Experiments

Dataset and Diseases

For our study, we selected a popular health forum *Medicitalia.it* (presumably the most famous and consulted medical website in Italy) from which to recover conversations related to the selected disease d (<https://www.medicalita.it/consulti/?pagina=1>). It offers medical content, a forum in which users interact with each other and a section dedicated to online medical consultations. This section includes to date more than half a million consultations with answers from at least one doctor. To create a collection of patient-doctor

conversations, we developed a software module for scraping the selected web forum (Note that, due to current *GDPR* regulations on privacy, patient names are not available, nor it is possible to link messages of the same patient in different threads. Furthermore, the text of messages cannot be published.).

To test the proposed framework, we selected 2 diseases and 2 tasks:

- $d1 = \text{diabetes}$, $t1 = \text{quality of life assessment}$
- $d2 = \text{hypothyroidism}$, $t2 = \text{therapy analytics of Levothyroxine}$ (the most common therapy for hypothyroid patients).

We considered diabetic and hypothyroid patients since they are two very common diseases that affect millions of patients around the world. After applying our classification procedure, we obtained 10,388 threads on topics concerning diabetes, and 27,125 thyroid-related threads.

Compared Systems

We compare our GTCACS topic detection techniques with two popular models:

- *Latent Dirichlet Allocation (LDA)* [34]: a generative probabilistic approach for collections of discrete data like text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics; while, in turn, each topic is modeled as an infinite mixture over an underlying set of clusters probabilities, that provide an explicit representation of a document.
- *Non-negative Matrix Factorization (NMF)* [35]: a matrix factorization method in which we constrain matrices to be non-negative. NMF represents a tool for analyzing high-dimensional data as it automatically extracts sparse and meaningful features from a series of non-negative data vectors. This technique has an extensive range of uses, from modeling

topics, to dimensionality reduction applications and signal processing.

Evaluation Strategy

To evaluate the quality of the extracted topics, being in an unsupervised context, we adopted the following strategy. Initially, for each extracted topic we considered only the n most relevant words, next we encoded each term in its respective embedding vector, and finally we calculated the quality scores for the obtained clusters based on various evaluation metrics. Word embedding represents a technique where singular words are represented as real-valued dense vectors capable of capturing context of a term in a document, semantic/syntactic similarities and relations with other words. Specifically, we exploited pre-trained word vectors with FastText [36] model for the Italian language (Continuous

Bag Of Words with position weights, size 300 and ngrams characters of length 5).

We selected three metrics widely used for the unsupervised evaluation of clustering and we propose a custom measure, called *quality score*, based on an embedding representation of words in clusters. Below we briefly describe the applied evaluation metrics:

- *quality score (qs)*: In our proposed measure, let $T = \{t_1, t_2, \dots, t_k\}$ be the set of all the k extracted topics, where $t_i = \{v[w_1], v[w_2], \dots, v[w_n]\}$ represents the i -th topic containing the embedding vectors of its top n words; let $centroid(t_i)$ be the vector that represents the multidimensional mean of the word vectors in topic t_i ; let $m \in \mathbb{N}$ and $combinations(S, m)$ be the set that contains all the combinations of size m from the set S , then:

$$\text{quality_score} = \frac{\text{mean}\{\text{intra_topic_sim}(t) | \forall t \in T\}}{\text{mean}\{\text{inter_topics_sim}(t', t'') | \forall (t', t'') \in combinations(T, 2)\}} \quad (4)$$

$$\text{intra_topic_sim}(t) = \text{mean}\{\cos_sim(v[w]', v[w]'') | \forall (v[w]', v[w]'') \in combinations(t, 2)\} \quad (5)$$

$$\text{inter_topics_sim}(t', t'') = \cos_sim(\text{centroid}(t'), \text{centroid}(t'')) \quad (6)$$

- *silhouette score (ss)* [37]: It represents the mean silhouette coefficient of all samples. The silhouette coefficient of the single i -th sample is calculated using the mean intra-cluster distance (a : mean distance between a sample and all other points in the same class) and the mean nearest-cluster distance (b : mean distance

between a sample and all other points in the next nearest cluster).

$$ss_i = \frac{(b - a)}{\max(a, b)}$$

For this measure the best value is 1 and the worst value is -1, while values near 0 indicate overlapping clusters.

- *calinski-harabasz score (chs)* [38]: It is also known as the Variance Ratio Criterion. This

measure is defined as the ratio of between-clusters dispersion (B) and of inter-cluster dispersion (W) for all clusters (dispersion represents the sum of distances squared). Let n the number pf samples and k the number of clusters:

$$chs = \frac{B}{W} \cdot \frac{n - k}{k - 1}$$

A higher Calinski-Harabasz score relates to a model with better defined clusters.

- *davies-bouldin score (dbs)* [39]: It represents the average “similarity” measure between each cluster C_i for $i = 1, \dots, k$ and its most similar cluster C_j . Where, “similarity” is defined as a measure $R_{i,j}$ that represents the ratio between the within-cluster distance (s_i : the average distance between each point of cluster i and the centroid of that cluster) and the between-cluster distance ($d_{i,j}$: the distance between cluster centroids and i and j). For this measure, lower values indicate a better partition, and zero is the lowest possible score.

$$R_{i,j} = \frac{s_i s_j}{d_{i,j}}$$

$$dbs = \frac{1}{k} \sum_{i=1}^k \max_{i \neq j} R_{i,j}$$

Table 1 shows the scores obtained for the thyroid ($d2$) domain. Scores are calculated using the

previously defined metrics for the three compared approaches, varying the selected number of topics and considering the top $n = 100$ words for each topic. In Table 1 bold numbers show the best absolute values for each of the four considered topic validity scores, and varying T . We observe that GTCACS achieves the best performance scores, based on all the selected metrics. Similar results are obtained in the diabetes domain (An extensive study on the social phenotype of diabetes patients has been published in [40].).

Topic Analysis and Subclustering

Starting from the computed scores, and after asking the clinicians to perform a manual analysis on the clusters obtained with GTCACS, we decided to consider $T = 7$ discussion topics in the thyroid domain. With a similar procedure, we selected $T = 14$ topics in the diabetes domain. The list of top 30 words for each topic and each domain can be retrieved from http://iim.di.uniroma1.it/source_spinger.html.

We manually labeled the 14 diabetes topics as follows: *glycemia, hepatology/gastroenterology, gynecology, angiology, cardiovascular complications, neurology, ophthalmology/orthopedics, andrology, endocrinology, symptoms/feelings, dietary issues, clinical tests and drugs, diabetes in general, various problems*. The 7 thyroid-related topics were labelled as

Table 1 Scores obtained in the thyroid domain for the extracted topics (varying T), based on different metrics: qs “quality score,” ss “silhouette score,” chs “calinskiharabasz score,” dbs “davies-bouldin score”

	#T = 4	#T = 5	#T = 6	#T = 7	#T = 8	#T = 9	#T = 10
NMF	qs: 0.2534 ss: 0.0039 chs: 3.1494 dbs: 8.2332	qs: 0.2570 ss: 0.0019 chs: 3.5533 dbs: 8.0652	qs: 0.2473 ss: chs: 3.1568 dbs: 8.5802	qs: 0.2558 ss: chs: 3.5017 dbs: 8.5468	qs: 0.2480 ss: chs: 3.3287 dbs: 9.2956	qs: 0.2445 ss: chs: 3.3567 dbs: 9.4437	qs: 0.2423 ss: chs: 3.3180 dbs: 9.8470
LDA	qs: 0.2584 ss: 0.0046 chs: 3.2986 dbs: 7.8296	qs: 0.2716 ss: 0.0010 chs: 4.1939 dbs: 7.2232	qs: 0.2830 ss: chs: 4.6643 dbs: 7.4218	qs: 0.2818 ss: chs: 4.6028 dbs: 7.8375	qs: 0.2780 ss: chs: 4.7624 dbs: 7.5386	qs: 0.2750 ss: -0.0111 chs: 4.2769 dbs: 7.8888	qs: 0.2635 ss: chs: 3.9174 dbs: 8.3786
GTCACS	qs: 0.2963 ss: 0.0207 chs: 4.6084 dbs: 5.9530	qs: 0.2914 ss: 0.0119 chs: 4.4443 dbs: 6.7542	qs: 0.2975 ss: 0.0102 chs: 4.7674 dbs: 6.2792	qs: 0.3020 ss: 0.0039 chs: 5.1189 dbs: 6.8415	qs: 0.2909 ss: 0.0006 chs: 4.7867 dbs: 7.2019	qs: 0.2993 ss: 0.0044 chs: 5.0705 dbs: 6.8437	qs: 0.2800 ss: chs: 4.2943 dbs: 8.3366

follows: *gynecology, thyroid nodules, cardiovascular complications, pregnancy, clinical tests and drugs, thyroid in general, symptoms/feelings*. As expected, there are similarities in the two sets of topics, while differences are related to the specificity of each disease. Our detected topics are particularly useful to cluster patients messages according to different dimensions and analysis objectives:

1. *Dimension 1: comorbidities.* First, we note that the first nine diabetes topics and the first four hypothyroidism topics are useful to stratify patients according to their reported comorbidities (note that hypo/hyper glycemia is a syndrome, not a disease, and pregnancy is a condition. We are aware that the term “comorbidity” is abused here.). These subsets of topics represent the first detected dimension of our study. Analyzing messages of patients affected by a given disease grouped by their comorbidities and conditions can be greatly helpful for doctors. For example, this message refers to the elder diabetic mother of the message author, with several comorbidities: “*my 78 year old mother suffers from atherosclerotic vasculopathy from arterial hypertension and insulin dependent diabetes mellitus. For some months now she has completely lost her balance and often falls.*”
2. *Dimension 2: symptoms.* The topic “symptoms” includes a variety of physical conditions that would require a better subcategorization for the purpose of therapy analytics. To this end, we repeated the procedure described in section “[Topic Modeling](#)” and found relevant

subclusters of words representing groups of disease-specific symptoms. For example, in the thyroid domain, we found five subclusters of adverse drug reactions to the treatment with Levothyroxine tablets, shown in Table 2 (only top frequency words). Again, this is extremely useful to group patient messages according to their symptoms, to detect adverse drug reactions: e.g.: “*I have autoimmune Hashimoto’s thyroiditis and I regularly take eutirox 75[...]. Since last May I have been accusing various problems that have culminated in the last period with episodes of vomiting, nausea...*”

3. *Dimension 3: emotions.* Finally, with the purpose of evaluating the feelings and emotions of these on-line patients, we further added six clusters of words identifying primary emotions, based on the work in [41, 42], and using the sentiment keywords presented in [43]. An excerpt of these emotion words is shown in Table 3 (10 words) (The complete list is shown in the previously provided link). Grouping messages along this dimension may foster a better understanding of patients’ feelings and quality of life: “*For 6 years I have been taking 75mg eutirox daily (since December 2014 50mg every day) because of thyroid nodules [...] in the last period I feel depressed and as if I was fainting in the heart, I feel like a strong blow in the heart and then the nothing that leaves me with strong shivers of fear.*”

These three dimensions (hereafter D1, D2 and D3), are at the basis of the social phenotype analytics phase, presented in section “[Data Analytics](#).”

Table 2 Subclusters of symptoms for hypothyroid patients (top ranked words)

Symptom	Terms
Cardiovascular disorders	<i>Tachycardia, palpitations, arrhythmia, heart attack, myocardium, heart, cardiac, hypertension, hyper-tense</i>
Nervous system disorders	<i>Headache, tremor, trembling, restless, restlessness, sleepless, insomnia, agitation, agitated, nervousness, nervous, anxiety, anxious</i>
Gastrointestinal disorders	<i>Nausea, diarrhea, vomiting, weight loss, appetite</i>
Systemic disorders	<i>Redness, fever, sweating, hyperhidrosis, flushing</i>
Reproductive system disorders	<i>Menstrual, cycle, menstruation</i>

Table 3 Emotion clusters (first 10 words in alphabetic order), based on the keywords in [43]

Emotion	Terms
Happiness	<i>Cheerful, cheerfulness, complacency, contented, contentment, delighted, enthusiasm, enthusiastic, euphoria, euphoric</i>
Sadness	<i>Afflicted, anguish,anguished, bitterness, bored, boredom, boring, depressed, depression, desolate</i>
Anger	<i>Agitated, agitation, anger, angry, annoyed, exasperated, exasperation, extremely angry, frustrated, frustration</i>
Disgust	<i>Abhorrent, abominable, contempt, degradation, degrading, disgust, disgusted, disgusting, horrendous, horrible</i>
Surprise	<i>Amazed, amazement, appalled, bewildered, confused, confusion, discovery, dismay, disoriented, doubt</i>
Fear	<i>Afraid, agitated, agitation, alarmed, anguish, anguished, anxiety, anxious, apprehension, breading</i>

Data Analytics

Data analytics in healthcare is evolving into a promising field that takes advantage of the increasing daily availability of health information to extract insights for producing better-informed decisions [44]. By discovering associations, understanding patterns and examining trends within patient-related health data, data analytics applications offers the potential to improve the quality of care at more moderate costs and with better outcomes. The misalignment between patients and doctors focuses remains a recognized problem in medical literature. In light of this, the information analyzed in this section is clearly relevant because it is provided anonymously by patients, according to their perceived priorities, rather than in response to specific questions formulated by specialists.

As anticipated in section “[Introduction](#),” in this study we are specifically focused on analyzing in detail the feelings, adverse drug reactions, symptoms, and everyday lifestyle of diabetic and hypothyroid patients.

Towards this objective, we performed various analyses to characterize messages from patients affected by our two considered diseases along the three dimensions identified in section “[Evaluation and Experiments](#).[“](#) This type of analysis is useful to help doctors understand in a more nuanced way the patients’ feelings. Thanks to our proposed methodology, doctors can easily identify specific and statistically relevant subsets of automatically grouped messages (for example, messages of hypothyroid patients with cardiological complications, expressing emotions such as fear or sadness).

Hereafter we provide examples of data analytic that can be carried out using our proposed methodology. For the sake of space, we do not carry out a detailed analysis of both diseases. The interested reader can refer to [40].

We first study the relevance of primary emotions (shown in previous Table 3) within a corpus of d -related messages. For each cluster of words associated with a primary emotion, initially, we calculate the cosine similarities between the centroid of the current sentiment group and all the centroids of the messages; after that, we considered the average of these similarity values as the final mean score associated with the current emotion. Thus, at the end of the procedure, we are able to quantify the *significance* level of emotions for d -patients. As an example, the bar graph of Fig. 3, shows the relevance of primary emotions in hypothyroid patients undergoing standard Levothyroxine therapy. As we can see from this chart, the prevailing emotions are fear and sadness (an example is: “*I have read that tyrosint can cause agitation . . . could this have affected the worsening of symptoms? Since I take it I always feel nervous and agitated, I am afraid that this agitation will ruin my stomach . . . I always drink chamomile and once I tried with Valerian, but I can not calm down!*”).

Next, we can analyze the *correlation strength* between the six primary emotions (D3) and the patients’ sub categories (D1) identified in section “[Topic Analysis and Subclustering](#).[“](#) The result is shown in the heatmap of Fig. 4, always referring to hypothyroid patients. To examine the correlation, in order not to penalize less frequent topics, we adopted the Normalized Pointwise Mutual Information (NPMI) [45] as association measure.

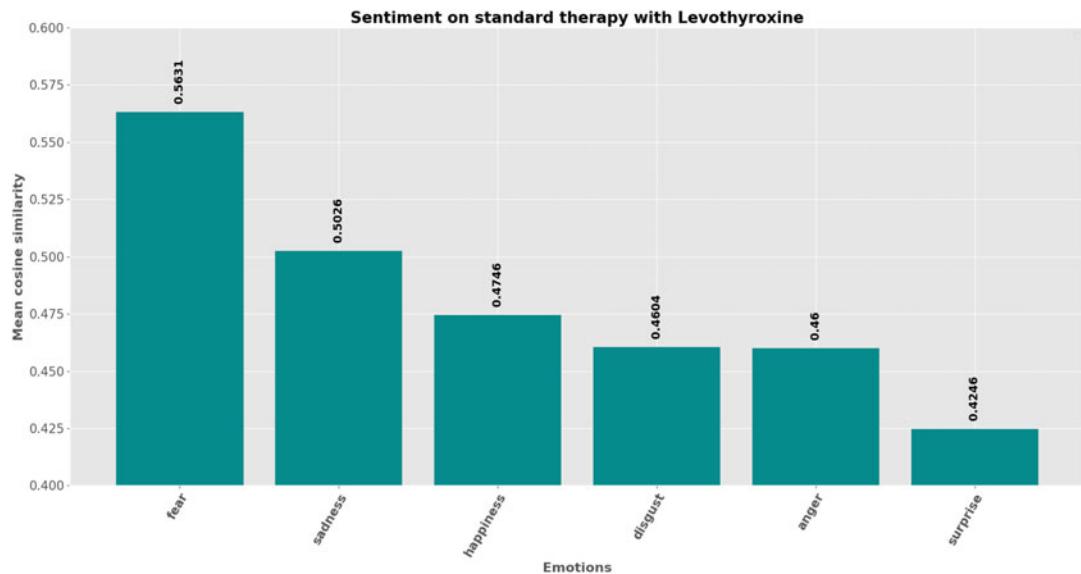


Fig. 3 Relevance of primary emotions (D3) in patients with hypothyroidism subject to standard therapy with Levothyroxine

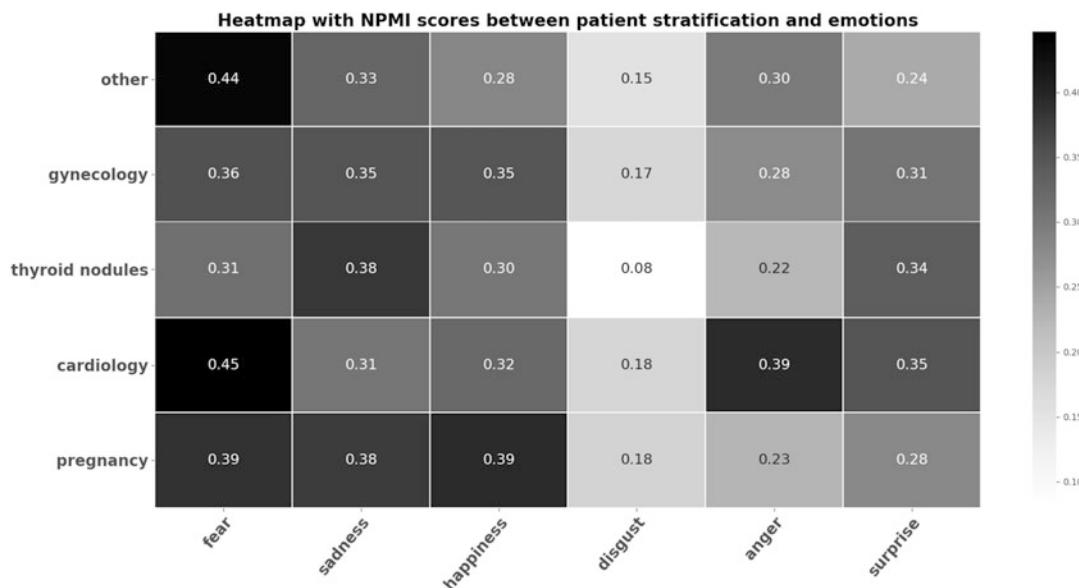


Fig. 4 Correlation between patient stratification (D1) and primary emotions (D3), for hypothyroid patients

Specifically, given two clusters of words $C1$ and $C2$, to calculate the NPMI score between these two groups: in the first place, we considered all the pairs of terms belonging to the Cartesian product between $C1$ and $C2$; next we calculated the NPMI scores for each of the obtained pairs of

words; finally, we selected the ten highest computed scores, and we considered the average of these top ten values as the final NPMI score of the two clusters $C1$ and $C2$. The heatmap shows that the highest NPMI value is between fear (as we already noticed, the most prevalent emotion among

our emotional topics) and the dimension “cardiology.” Not necessarily cardiovascular complications are the most severe, but the data show that they are the ones that patients fear the most.

Next, we can investigate ADRs in a *d*-related corpus of messages. In the hypothyroid/Levothyroxine sub collection, we can use the five symptom clusters (D2) shown in previous Table 2. The bar chart of Fig. 5, calculated as for in Fig. 3, shows that the most relevant groups of complained symptoms for hypothyroid patients appear to be connected to cardiovascular disorders and nervous system problems, such as depression.

Furthermore, we can examine the correlations between the symptom clusters (D2) and the patient stratification (D1). In Table 6 we apply this analysis to diabetic patients. The heatmap of Fig. 6 shows some obvious correlation between comorbidities and symptoms (e.g., ophthalmology/orthopedics and eye problems, or andrology and sexual disorders), but also other interesting correlations, like polyuria symptoms for hepatology/gastroenterology-related comorbidities, and again the widespread relevance of cardiovascular and dietary (polyphagia and weight) problems. An example

of symptoms reported by (the son of) an elder diabetes patient with cardiological comorbidities is: “*My father suffering from diabetes has attacks of angina, now it happens even when at rest. He always has a feeling of heaviness in the stomach and pain in the arm shoulder, that is alleviated with the Cavasini, but then it returns to him often times at night [...] We do not know if his treatment is too heavy for him, for his age, because since he has been doing it he says he feels worse . . .*”

Finally, in Fig. 7, we show the correlations between the symptom clusters (D2) and the primary emotions clusters (D3), again for the case of hypothyroid patients. We observe that patients communicate more openly their emotions when reporting depression (nervous system disorders) and cardiovascular complications.

To summarize the results of our analysis, we can say that psychological and cardiovascular complications represent the types of ADR with the highest association with hypothyroid patients treated with Levothyroxine. This fact reveals how psychological aspect has a key role and reinforces the evidence of dissatisfaction and anxiety of this type of patients. For diabetic patients, the most

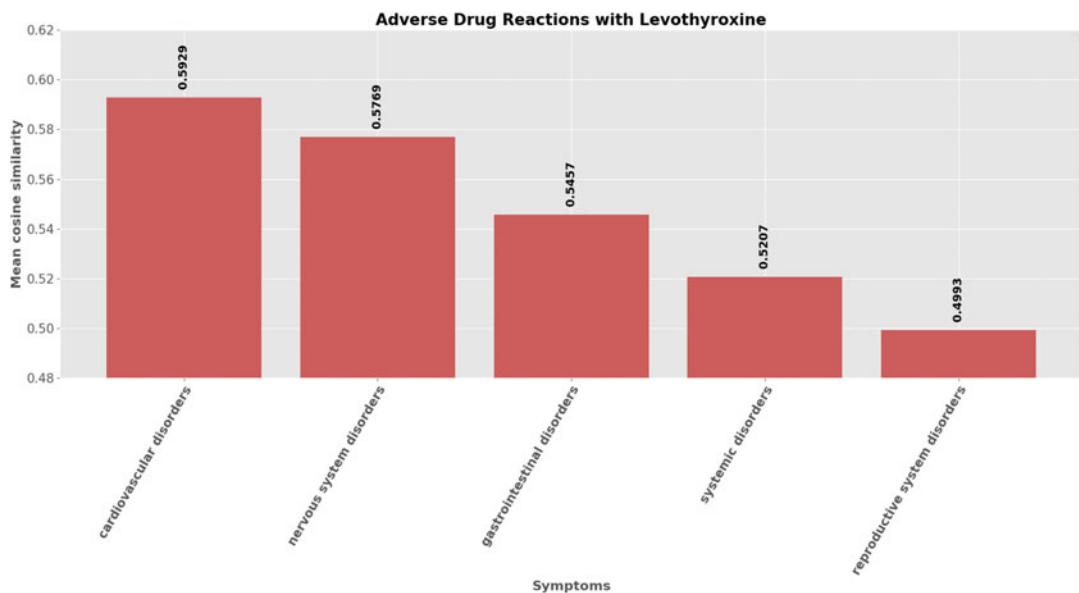


Fig. 5 Adverse Drug Reactions for hypothyroid patients: mean cosine similarity between the centroid of symptom clusters (D2) and the centroid of documents (messages)

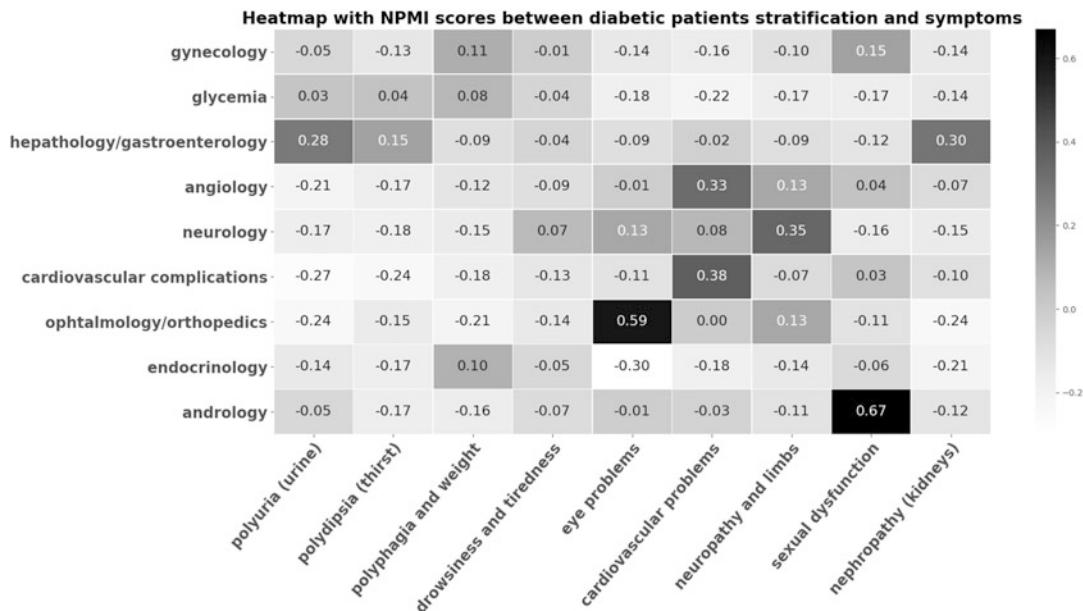


Fig. 6 Correlation between patient stratification (D1) and symptoms (D2) for diabetic patients

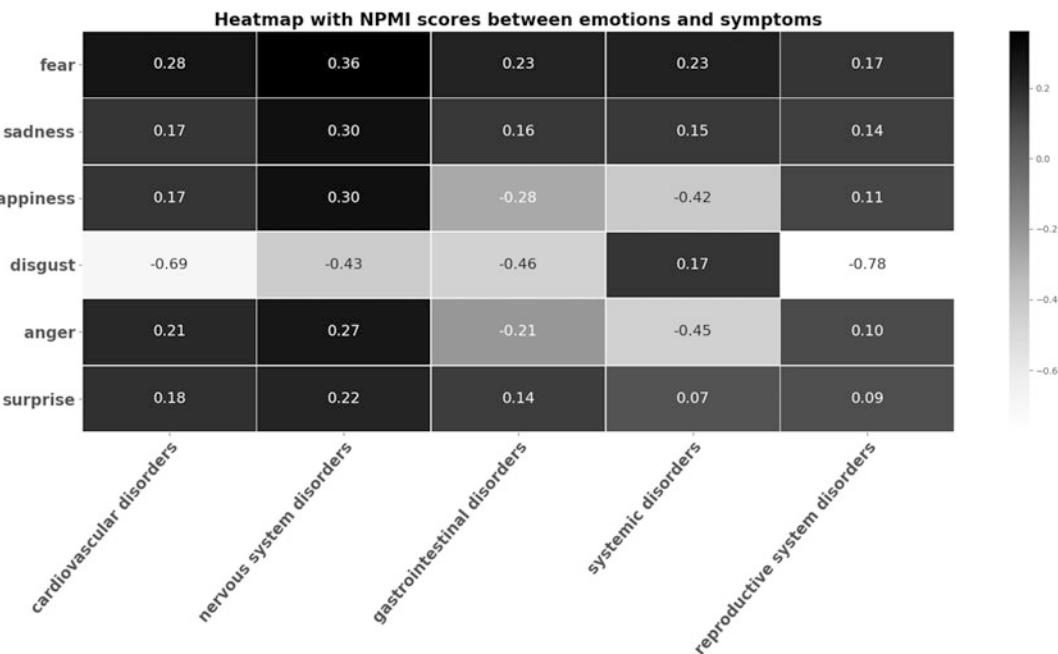


Fig. 7 Correlation between primary emotions (D3) and symptoms (D2) for hypothyroid patients

frequently reported problems concern glycemic complications (more in detail, insulin-related problems) and dietary problems. Although these

problems are not considered by diabetologists as representing a severe limitation of a patient's life (Except for the assumption of insulin, which is

known to be a major cause of discomfort.), at least in comparison with other symptoms affecting eyes, limbs and kidneys [46], our analysis shows that indeed they are reported by the patients in forums as being a major cause of discomfort.

Conclusion

In this chapter we presented a framework for extracting and analyzing the *social phenotype* of a disease, that we define as the systematic detection, analysis and categorization of messages generated by patients affected by a disease. Data on these patients can be collected in the large from health forums, where users more freely present to doctors a subjective and unfiltered picture of their health conditions, lifestyle and emotional status. Fine-grained, large scale categorization of patients' messages is supported by the detection of discussion topics, that we subsequently organized along three dimensions: comorbidities, symptoms, and emotions.

For an effective categorization of patients' messages, we designed a data processing workflow based on:

1. an AutoML pipeline, to accurately classify messages of patients affected by a selected disease;
2. an innovative topic extraction methodology, Generative Text Compression with Agglomerative Clustering Summarization (GTCACS), to identify different dimensions along which patients can be further stratified, in support of therapy analytics.

The framework has been applied to the analysis of the social phenotype of diabetic patients and hypothyroid patients treated with Levothyroxine, extracted from a popular health forum in Italy. We have shown that this type of analysis is useful to help doctors understand in a more nuanced way patients' feelings and types of complained ADRs, which are often overlooked during ordinary patient-doctor communications. Thanks to our proposed therapy analytic framework, doctors can analyze specific and statistically relevant

subsets of automatically grouped messages, rather than inspecting the entire unordered collection.

References

1. Gallan AS. Evaluation and measurement of patient experience. *Patient Exp J.* 2014;1:5.
2. Stegemann S, Ternik RL, Onder G, Khan M, van Riet-Nales D. Defining patient centric pharmaceutical drug product design. *AAPS J.* 2016;18:1047.
3. McCabe R, PGT H. Miscommunication in doctor-patient communication. *Top Cogn Sci.* 2018;10:409–24.
4. Cunillera O, Tresserras R, Rajmil L, Vilagut G, Brugulat P, Herdman M, et al. Discriminative capacity of the EQ-5D, SF-6D, and SF-12 as measures of health status in population health survey. *Qual Life Res Int J Qual Life Asp Treat Care Rehab.* 2010;19:853–64.
5. Schegloff E. When 'others' initiate repair. *Appl Linguis.* 2000;06:21.
6. Smailhodzic E, Hooijmans W, Boonstra A, Langley DJ. Social media use in healthcare: a systematic review of effects on patients and on their relationship with healthcare professionals. *BMC Health Serv Res.* 2016;16(1):442.
7. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Harnessing the cloud of patient experience: using social media to detect poor quality healthcare. *BMJ Qual Saf.* 2013;22:251.
8. Velardi P, Stilo G, Tozzi AE, Gesualdo F. Twitter mining for fine-grained syndromic surveillance. *Artif Intell Med.* 2014;61(3):153.
9. Oconnor K, Pimpalkhute P, Nikfarjam A, Ginn R, Smith K, Gonzalez G. Pharmacovigilance on Twitter? Mining tweets for adverse drug reactions. *AMIA Ann Symp Proc/AMIA Symp.* 2014;2014:924–33.
10. Rains SA, Keating D. The social dimension of blogging about health: health blogging, social support, and well-being. *Commun Monogr.* 2011;78:511–34.
11. Vydiswaran VGV, Liu Y, Zheng K, Hanauer DA, Mei Q. User-created groups in health forums: what makes them special? In: Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, 1–4 June 2014; 2014.
12. Miller E, Pole A. Diagnosis blog: checking up on health blogs in the blogosphere. *Am J Public Health.* 2010;100(8):1514–9.
13. Wilson E, Kenny A, Dickson-Swift V. Using blogs as a qualitative health research tool: a scoping review. *Int J Qual Methods.* 2015;01:12.
14. van Eenbergen MC, van de Poll-Franse LV, Krahmer E, Verberne S, Mols F. Analysis of content shared in online cancer communities: systematic review. *JMIR Cancer.* 2018;4(1):e6.
15. Paul MJ, Sarker A, Brownstein J, Nikfarjam A, Scotch M, Smith K, et al. Social media mining for

- public health monitoring and surveillance. Pacific Symposium on Biocomputing. World Scientific Publishing Co. Pte Ltd. 2016;21:468–79.
- 16. Bahja M, Lycett M. Identifying patient experience from online resources via sentiment analysis and topic modelling. In: Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies. BDCAT '16; 2016. p. 94–9.
 - 17. Della Rosa S, Sen F. Health topics on Facebook groups: content analysis of posts in multiple sclerosis communities. *Interact J Med Res.* 2019;8(1):e10146.
 - 18. Correia RB, Wood IB, Bollen J, Rocha LM. Mining social media data for biomedical signals and health-related behavior. *CoRR.* 2020;abs/2001.10285. Available from: <https://arxiv.org/abs/2001.10285>
 - 19. Hamed AA, Wu X, Erickson R, Fandy T, Twitter K-H networks in action: advancing biomedical literature for drug search. *J Biomed Inform.* 2015;56:157–68.
 - 20. Belousov M, Milosevic N, Dixon WG, Nenadic G. Extracting adverse drug reactions and their context using sequence labelling ensembles in TAC2017. In: Proceedings of the 2017 Text Analysis Conference, TAC 2017, Gaithersburg, Maryland, 13–14 November 2017, NIST; 2017.
 - 21. Yang CC, Yang H, Jiang L, Zhang M. Social media mining for drug safety signal detection. In: Yang CC, Chen H, Waclaw HD, Combi C, Tang X, editors. Proceedings of the 2012 International Workshop on Smart Health and Wellbeing, SHB 2012, 29 October 2012, Maui, ACM; 2012. p. 33–40.
 - 22. Alhuzali H, Ananiadou S. Improving classification of adverse drug reactions through using sentiment analysis and transfer learning. In: Proceedings of the 18th BioNLP workshop and shared task. Florence: Association for Computational Linguistics; 2019. p. 339–47. Available from: <https://www.aclweb.org/anthology/W19-5036>
 - 23. Nguyen T, Larsen ME, O'Dea B, Phung DQ, Venkatesh S, Christensen H. Estimation of the prevalence of adverse drug reactions from social media. *Int J Med Inform.* 2017;102:130–7.
 - 24. Nikfarjam A, Sarker A, O'Connor K, Ginn RE, Gonzalez-Hernandez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *J Am Med Inform Assoc.* 2015;22(3):671–81.
 - 25. Huynh T, He Y, Willis A, Rueger S. Adverse drug reaction classification with deep neural networks. In: Calzolari N, Matsumoto Y, Prasad R, editors. COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 11–16 December 2016, Osaka. ACL; 2016. p. 877–887.
 - 26. Stilo G, Vincenzi MD, Tozzi AE, Velardi P. Automated learning of everyday patients' language for medical blogs analytics. In: Angelova G, Bontcheva K, Mitkov R, editors. Recent Advances in Natural Language Processing, RANLP 2013, 9–11 September 2013, Hissar, Bulgaria. RANLP 2013 Organising Committee/ACL; 2013. p. 640–8.
 - 27. Weng Z. From conventional machine learning to AutoML. *J Phys Conf Ser.* 2019;1207:012015.
 - 28. Hutter F, Hoos HH, Leyton-Brown K. Sequential model-based optimization for general algorithm configuration. In: Proceedings of the 5th International Conference on Learning and Intelligent Optimization. LION05. Berlin/Heidelberg: Springer; 2011. p. 507523. Available from: https://doi.org/10.1007/978-3-642-25566-3_40
 - 29. Zhao JJ, Mathieu M, LeCun Y. Energy-based Generative Adversarial Network. *ArXiv.* 2016;abs/1609.03126.
 - 30. Glover J. Modeling documents with Generative Adversarial Networks. *ArXiv.* 2016;abs/1612.09122.
 - 31. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Nets. In: NIPS; 2014.
 - 32. Yu Y, Gong Z, Zhong P, Shan J. Unsupervised representation learning with deep convolutional neural network for remote sensing images. In: Zhao Y, Kong X, Taubman D, editors. Image and graphics. Cham: Springer International Publishing; 2017. p. 97–108.
 - 33. Ward JH. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc.* 1963;58(301):236–44.
 - 34. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. *J Mach Learn Res.* 2003;3(null):9931022.
 - 35. Paatero P, Tapper U. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics.* 1994;5(2):111–26.
 - 36. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. *Trans Assoc Comput Linguis.* 2017;5:135–46. Available from: <https://www.aclweb.org/anthology/Q17-1010>
 - 37. Rousseeuw P, Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Comput Appl Math.* 20:53–65. Journal of Computational and Applied Mathematics. 1987 11;20:53–65
 - 38. Caliski T, Harabasz J. A dendrite method for cluster analysis. *Commun Statist.* 1974;3(1):1–27.
 - 39. Davies DL, Bouldin DW. A cluster separation measure. *IEEE Trans Pattern Anal Mach Intell.* 1979;PAMI-1(2):224–7.
 - 40. Lenzi A, Maranghi M, Stilo G, Velardi P. The social phenotype: extracting a patient-centered perspective of diabetes from health-related blogs. *Artif Intell Med.* 2019;101:101727. Available from: <https://doi.org/10.1016/j.artmed.2019>
 - 41. Oppong S, Asamoah D, Oppong E, Lamptey D. Business decision support system based on sentiment analysis. *Int J Inform Eng Electron Bus.* 2019;11:36–49.
 - 42. Ekman P. An argument for basic emotions. *Cognit Emot.* 1992;6(34):169–200.

43. Chen Y, Huang CR, Lee S. Automatic recognition of emotion based on a cognitively motivated emotion annotation system. *J Cogn Sci.* 2011;12:279–96.
44. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inform Sci Syst.* 2014;2:3.
45. Bouma G. Normalized (Pointwise) Mutual Information in Collocation Extraction. Proceedings of the Biennial GSCL Conference 2009. 2009 01.
46. Clarke NG, Fox K, Grandy S. Symptoms of diabetes and their association with the risk and presence of diabetes. *Diabetes Care.* 2008;30:2868–74.



AIM and Transdermal Optical Imaging 82

Andrew Barszczyk, Weihong Zhou, and Kang Lee

Contents

Introduction	1144
Characterizing the Cardiovascular System: Benefits, Obstacles, and Breakthroughs	1144
Transdermal Optical Imaging	1146
Overcoming Measurement Obstacles with Transdermal Optical Imaging Technology	1146
Scientific Foundations of Transdermal Optical Imaging	1146
Present Advances Using Transdermal Optical Imaging	1150
Trends in Medicine and the Potential Impact of TOI	1151
Future Uses and Challenges for TOI	1153
References	1155

Abstract

Cardiovascular parameters like blood pressure, heart rate, heart rhythm, and heart rate variability are highly useful in assessing patient health,

disease risk, and response to treatment. However, technological limitations curtail their measurement in many cases. The recent development of transdermal optical imaging (TOI) technology has made it possible to extract high-quality blood flow information from conventional video of a patient's face and then use it to accurately estimate these cardiovascular parameters. TOI technology could thus be implemented on any device capable of capturing and processing video (e.g., any modern smartphone) and thus constitute a comfortable, convenient, and ubiquitous tool for measuring cardiovascular parameters.

TOI builds upon remote photoplethysmography in part by using machine learning to extract robust blood flow information from video of the face. This signal can be used directly to compute heart rate, heart

Supplementary Information: The online version of this chapter (https://doi.org/10.1007/978-3-030-64573-1_250) contains supplementary material, which is available to authorized users.

A. Barszczyk · W. Zhou (✉)
Health Management Centre, Drum Tower Hospital
Affiliated to Nanjing University Medical School, Nanjing,
China
e-mail: andrewbarszczyk@nuralogix.ai

K. Lee (✉)
Dr. Eric Jackman Institute of Child Study, University of
Toronto, Toronto, ON, Canada
e-mail: kang.lee@utoronto.ca

rhythm, and heart rate variability. Information from signal features containing blood pressure information has been combined with the help of machine learning to accurately estimate systolic and diastolic pressures. Further development of this technology is likely to enable the assessment of additional physiological parameters (e.g., respiration rate, SpO₂), disease risks (e.g., hypertension, diabetes), blood biomarker concentrations (e.g., cholesterol, HbA1c), and even mental health conditions (e.g., depression, anxiety).

With the necessary regulatory approvals and clinical trials, TOI-based tools would enable accurate, convenient, and contactless screening, diagnosis, and monitoring of patient health. They would revolutionize healthcare delivery through better access and efficiency and thus not only reduce costs but also improve health worldwide.

Keywords

Transdermal optical imaging · Smartphone · Blood pressure · Telemedicine · Remote health monitoring · Heart rate variability · Heart rate · Heart rhythm

Introduction

See Video 1.

Characterizing the Cardiovascular System: Benefits, Obstacles, and Breakthroughs

Cardiovascular parameters like blood pressure, heart rate, and heart rhythm are some of the most vital measures in clinical health assessment. Heart rate variability (the subtle variation in timing between one heartbeat and the next) provides insight into cardiovascular health and general wellness. Such measures are routinely employed in the medical setting and increasingly at home and elsewhere. While the benefits of comprehensive monitoring are becoming increasingly

apparent, so are the technological barriers preventing cardiovascular monitoring from being used to its full potential.

Blood Pressure Measurement

Hypertension Screening, Diagnosis, and Management

Hypertension (high blood pressure) is a major modifiable risk factor for cardiovascular disease, with a global prevalence of over 30% [1]. There is an immense need to employ blood pressure measurement in screening, diagnosing, and managing hypertensive blood pressures down to target levels. Despite the importance of these activities, hypertension awareness, treatment, and control rates remain less-than-ideal in high-income countries (67%, 56%, and 28%, respectively) and very limited in middle- and low-income countries (38%, 29%, and 8%, respectively) [1, 2].

Awareness is limited by inadequate screening in the clinic [3], and the requirement for special equipment to measure blood pressure is an obstacle to opportunistic self-screening at home. Diagnosis proceeds according to algorithms that most often require home or ambulatory measurements in addition to clinic measurements, and the management of high blood pressure via lifestyle modification or medication calls for regular blood pressure measurement on an ongoing basis at home [4]. The daily home (“out-of-office”) monitoring recommendation is to take two measurements in the morning and two in the evening with a cuff-based oscillometric device. However, such measurements remain uncomfortable, and the need for multiple measurements (to average out naturally occurring oscillations in blood pressure) is time-consuming. Such factors might reduce patients’ willingness to monitor their blood pressure and hinder blood pressure control. Further, comprehensive blood pressure profiles are increasingly valued in characterizing hypertension [5]. However, the bulkiness cuff-based devices make them inconvenient to bring along outside of the home, thus limiting the ability to comprehensively monitor one’s blood pressure over a range of daily activities (e.g., at work).

General Health Assessment

Blood pressure measurement is a primary tool in assessing patient health in the clinic. For instance, high blood pressure could indicate cardiovascular compromise (stroke or myocardial infarction), and low blood pressure could indicate the presence of shock (septic, cardiogenic, anaphylactic). Measurements are conducted by auscultation with a sphygmomanometer or by automated oscillometric devices. A limitation of these cuff-based tools is that they are potentially unsanitary, and they require staff to get close to patients (potentially increasing the likelihood of infectious disease transmission). Further, cuff-based tools cannot be applied remotely in telemedicine consultations, thus precluding a comprehensive health assessment for those without devices at home.

General Health Monitoring

Blood pressure is also monitored continuously and semicontinuously in the clinic. For instance, patients are monitored perioperatively to track depth of anesthesia and both perioperatively and postoperatively for complications (e.g., blood vessel rupture or sepsis) [6]. Continuous monitoring is conducted most accurately (albeit invasively) by arterial line. Noninvasive continuous alternatives include finger-based devices (calibrated by brachial cuff), although such devices are very costly. A frequent alternative to these invasive or costly methods when possible is semi-continuous monitoring via auscultation or oscillometric device. However, their disadvantage is that nurses must occupy more of their time taking measurements. It has been proposed that some patients monitored semicontinuously might be able to spend less time in hospital by self-monitoring at home after minor surgical procedures. However, the logistical hurdle of distributing (and later collecting) cuff-based devices could limit the practicality of this option.

Heart failure patients require regular out-of-office monitoring to ensure that their blood is being pumped adequately [7]. The comprehensiveness of the monitoring required once again raises concerns about the comfort and convenience of cuff-based measurements.

Heart Rate, Heart Rhythm, and Heart Rate Variability

General Assessment and Monitoring

Heart rate is an easily accessible clinical indicator of various conditions that may require further assessment. Abnormally fast heart rates (tachycardia, >100 bpm) could indicate dehydration, infection, or cardiogenic shock. Abnormally slow heart rates (bradycardia, <60 bpm) could indicate depressant effects from medication or that something is wrong with the pacing mechanism of the heart. Heart rate can be crudely assessed by palpation (requiring time and contact with the patient), but in the clinic it is most often measured by various specialized equipment (e.g., pulse oximeter, blood pressure monitor, electrocardiogram (ECG), or auscultation by a trained professional).

Heartbeat rhythm and regularity are also important health indicators. Gross irregularities in inter-beat intervals (e.g., complexes of premature ventricular contractions) constitute arrhythmias that could suggest greater risk of disease or require treatment. Conversely, when it comes to the subtle timing differences between heartbeats (termed “heart rate variability”), more variability implies greater responsiveness to blood pressure and respiratory fluctuations. Higher heart rate variability is prognostically favorable for various cardiovascular diseases [8]. Further, heart rate variability has recently been identified as one of the earliest physiological predictors of sepsis/infection [9]. Frequency analysis of heart rate variability can closely approximate the state of the autonomic nervous system (sympathetic-parasympathetic balance); increased spectral power at low frequencies relative to high frequencies suggests an increase in sympathetic tone [8]. Certain time domain measures of heart rate variability are also informative and are strongly associated with parasympathetic activity [10].

Heart rhythm irregularities can be detected by a trained professional using auscultation, but most often they are assessed by electrocardiography (ECG). Heart rate variability is most often measured by ECG. ECG is highly accurate but requires specialized equipment and the attachment of adhesive probes to the patient’s chest to

attain the highest-quality recordings. This requirement has largely limited such assessments to specialty care (e.g., cardiology, neurology) and made them impractical in large-scale screens, general health assessments, and remote consultations.

Stress Assessment

The appreciation that psychological stress has a major impact on health and general wellness is driving demand for more frequent stress assessment in more people. Psychological stress has been traditionally assessed through questionnaires. However, a growing literature has associated various permutations of heart rate variability with psychological stress, thus creating a quick, efficient, and objective method of stress assessment [11]. However, its potential is limited by the requirement for ECG equipment.

The Need for Technological Breakthrough

It is becoming apparent that more comprehensive monitoring of cardiovascular parameters could have immense benefits but that the inconvenience, discomfort, and inaccessibility of existing tools present obstacles to measurement. A technological breakthrough is needed to enable the creation of a new generation of tools that are more comfortable, more convenient, and more accessible to truly realize the full potential of maximizing health and wellness through the measurement of these parameters.

Transdermal Optical Imaging

Overcoming Measurement Obstacles with Transdermal Optical Imaging Technology

One solution that addresses all of these measurement obstacles is transdermal optical imaging (TOI) technology [12–17], which uses artificial intelligence to accurately measure cardiovascular parameters from video of a patient's face. The technology extracts a continuous pulsatile blood flow signal from video. This signal is rich in physiological information and interrelated

throughout the body [18]. As such, it can be used to estimate parameters like brachial (upper arm) blood pressure [14, 16, 17], heart rate [13], heart rhythm, and heart rate variability/stress [13].

TOI technology has a basis in conventional video and thus can be implemented as software on existing computing devices (e.g., smartphones, tablets, or any other device capable of capturing and processing video), thus transforming them into measurement tools without the need for special equipment. Such tools would be contactless and thus could be used more comfortably than many existing tools. Tools using TOI technology could be conveniently employed anytime and anywhere, with measurements carried out comfortably, safely (e.g., without contact), and efficiently.

Scientific Foundations of Transdermal Optical Imaging

Biomechanics and Video Capture

TOI technology is a novel variant of a more than decade-old technique called remote photoplethysmography [19]. With each beat of the heart, blood is circulated throughout the arterial system and produces a pressure pulse (a small expansion of the arteries) that propagates through the arterial system [20]. As this pulse passes through the superficial arteries of the skin, the arterial expansion compresses microvascular blood toward the surface of the skin. This puts additional blood into range of ambient light that has penetrated the superficial layers of the skin. The additional hemoglobin in this additional quantity of blood absorbs more of this light (according to its absorption spectrum), and consequently less light reemerges from the skin. Melanin in the skin also absorbs light, but its concentration remains constant from moment to moment. These cyclical attenuations of light can be captured by a consumer-grade video camera to reproduce the pressure pulses occurring within the artery. See Fig. 1 for schematic.

Distance from the camera, motion, skin tone, and variable lighting conditions can also affect what the camera captures, and so software

employing TOI technology addresses these issues at the video capture stage. Distance from the camera is approximately controlled by having the user place their head within a head-shaped outline on the graphical user interface, and a software-based face tracker will further warn the user if they are too close or too far away. Motion of the face relative to the camera is similarly tracked with software-based face detection; measurement will not begin, and ongoing measurements will be cancelled if too much motion is detected. Skin tone and variable lighting conditions are accounted for in part by calibrating camera exposure and white balance prior to the measurement. Measurements will not proceed, and ongoing measurements will be cancelled if there is insufficient lighting or too strong of a light gradient from one area of the face to the next. In this way, recording conditions are optimized and standardized.

Extracting Plethysmographic Signal from Video Using Artificial Intelligence

Some of the earliest variants of facial remote photoplethysmography averaged pixel intensity in the green channel and then filtered out noise in this raw signal with software-based digital signal filters [19, 21]. However, it was believed that much of this signal arose from mechanical (ballistocardiographic) movements of the head

(the subtle periodic movement of the head caused by blood pulsing up through the aorta and into the head) rather than color fluctuations caused by arterial pulsation under the skin surface. Principal components analysis (PCA) [22] was used to try to address this issue by identifying components of the signal that oscillated from one moment to the next relative to video data in the red channel (relatively little red light is absorbed by blood hemoglobin, and so it might be used as a baseline against which arterial pressure pulse oscillations can be detected) [23]. Another approach called the chrominance method [24, 25] isolated pulsatile signal in each color channel individually. A limitation of these methods is that they tend to identify periodicity based on the signal frequency with the highest signal power and therefore remain somewhat sensitive to contamination from periodic movements of the head [24]. Others have developed mathematical models to try to account for (and separate) the individual contributions of hemoglobin, melanin, and shadows in the extracted signal based on knowledge about their absorption spectra and camera spectral characteristics [26]. However, this deliberate modeling technique has had only limited success (perhaps due to the difficulty of adequately modeling the complexity of these factors).

TOI technology builds upon these approaches to achieve highly robust signal extraction

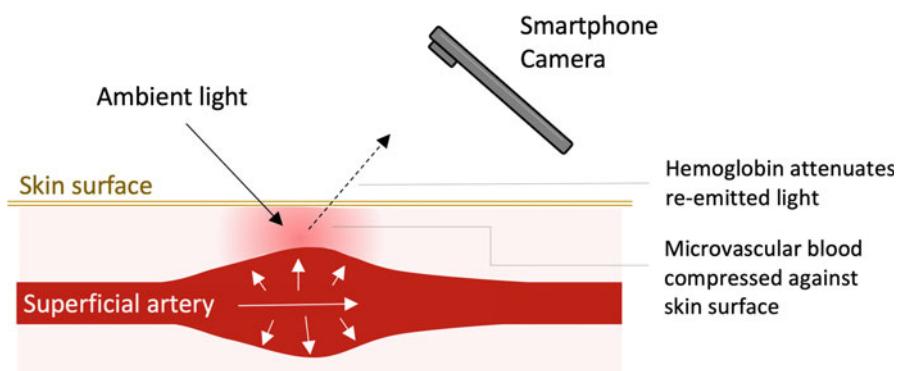


Fig. 1 Biomechanics and video capture for remote photoplethysmography. Pressure pulses created by the left ventricle of the heart arrive at the superficial arteries in the face. Here their expansion compresses microvascular blood toward the skin surface. The increased quantity of hemoglobin near the surface of the skin absorbs more

incident ambient light (the specific wavelengths corresponding to its absorption spectrum), and consequently less of this light reemits from the skin. These subtle attenuations of light are captured by a conventional video camera as subtle skin color fluctuations

[12]. Like other methods, TOI tracks specific regions of the face and extracts information from the red, green, and blue color channels (using multiple colors helps create a more robust signal) (Fig. 2a). But unlike any other approach, TOI uses a computational model trained using advanced machine learning techniques to determine the relation between skin color changes and arterial pressure pulse oscillations from one moment to the next. TOI encodes color information in each pixel as multiple layers (or “bitplanes”) of binary values (0 or 1) for each of the color channel (red, green, and blue), and this serves as input for the computational model. The model (previously trained to predict a continuous arterial pressure) then selects the bitplanes that best represent hemoglobin fluctuations to comprise the signal. This technique results in a high proportion of signal relative to noise (e.g., from variable lighting conditions or skin tone differences). A machine learning approach like this may account for additional complexity that is difficult to model with more deliberate techniques.

TOI further tracks multiple unique regions of interest (ROI) on the face (Fig. 2b). The reason for doing so is that different regions exhibit different spatiotemporal properties relative to ECG, despite all the pulses originating from the heart [12]. This is due to slight anatomical differences in facial vasculature, as well as the differential innervation of specific regions by either sympathetic or parasympathetic vasomotor neurons. For instance, sympathetic activity can strongly constrict subcutaneous blood vessels in the nose and lips and actively dilate vessels elsewhere like the forehead, cheeks, and chin [27]. Parasympathetic activity contributes to vasodilation in the lips and forehead. TOI thus captures rich information about the state of the vasculature and the autonomic nervous system that would be lost if signal from multiple regions of the face was simply averaged together. Finally, differences in temporal dynamics demonstrate that TOI signal is indeed driven by color changes rather than ballistic movements of the head [12].

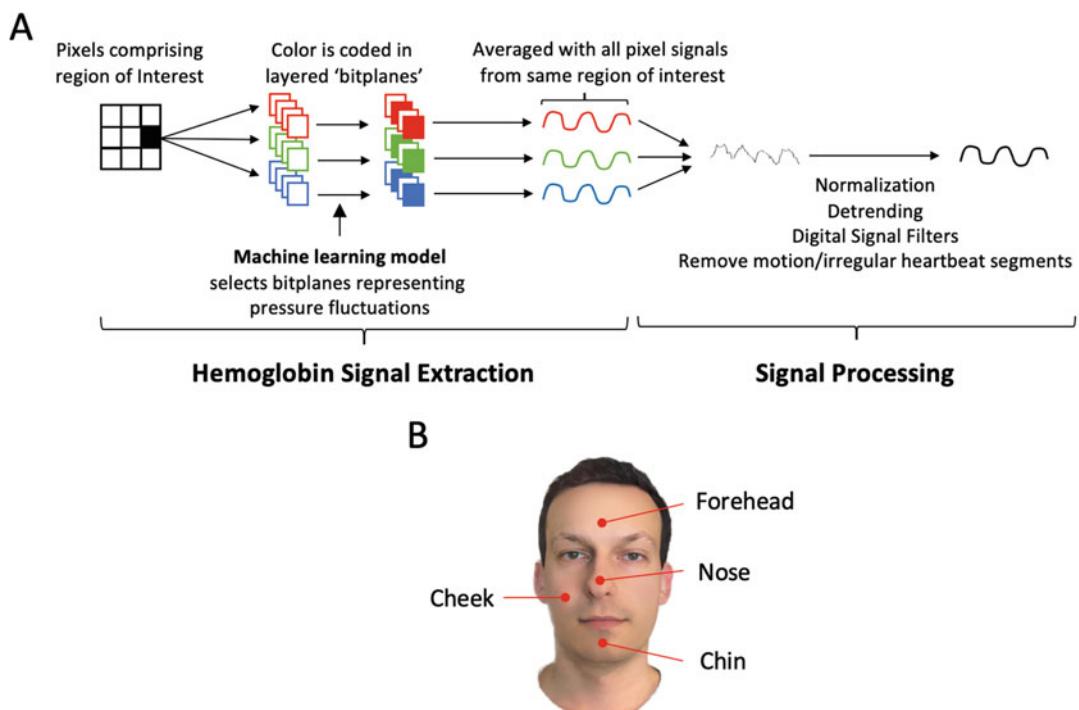


Fig. 2 Schematic of hemoglobin signal extraction and signal processing process (a) and major regions of interest (b)

After extracting raw signal in the red, green, and blue color channels, the signal is combined into a single raw blood flow signal (pressure pulse oscillation) for each ROI according to a special function [14]. During this process, each signal is normalized and detrended to account for variable lighting conditions. Digital signal filters (high, low, band-pass) remove low-frequency physiological oscillations (e.g., Mayer waves) and high-frequency noise. Signal elements suggesting motion and irregular heartbeats are also detected at this stage, and such segments are removed from the signal [17]. The final signal for each discrete region of the face is passed along to a variety of functions that extract information about physiological characteristics [14]. Some raw color information is passed along as well.

Cardiovascular Parameters and Their Relation to Blood Flow Features

Pulse Rate

Pulse rate corresponds to the patient's heart rate and is calculated as the number of pulses that arrive at the facial vasculature (from the heart). Pulses are counted as the number of major peaks within the signal segment and can be expressed as a frequency in hertz or a rate per minute [14].

Pulse Rate Variability

Pulse rate variability corresponds to the patient's heart rate variability, which is a measure of the variability in the intervals between heartbeats. Pulse rate variability measured by TOI closely tracks heart rate variability measured by ECG [13]. Pulse rate variability is calculated by detecting each pulse within the signal segment and then applying one several time domain calculations or frequency analyses that characterize the timing between pulses. For instance, the time domain SDNN measure of heart rate variability is calculated as the standard deviation of the inter-beat intervals. A larger SDNN value indicates more heart rate variability.

Blood Pressure

Systolic and diastolic blood pressures can also be estimated using plethysmographic signal features.

While many such features have been identified (e.g., augmentation index, pulse area) [28], no single feature has yet been found to provide enough information to accurately estimate blood pressure. It has thus been necessary to combine information from a variety of features to yield accurate estimates.

In an example of accurate blood pressure estimation, Luo and colleagues (2019) estimated blood pressure from a comprehensive set of 126 TOI blood flow features and 29 non-blood flow features [14] (Fig. 3). The blood flow features in that study and in a subsequent study by Yang et al. (2020) [17] can be divided into five groups: pulse shape, pulse energy (rates of change in pulse shape), pulse transit time, pulse rate variability, and pulse rate. **Pulse shape** features consisted of distances between various landmarks in the waveform, areas of certain sections within the waveform, or the ratio of one of these measurements with respect to another. Such features were calculated as means, maxima/minima, or measures of spread (e.g., standard deviation) across all pulse waveforms in the signal. **Pulse energy** features captured the rate at which certain pulse shape features emerged in the waveform as a function of time and were calculated as derivatives of pulse shape features. **Pulse transit time** is inversely related to the speed at which the pulse propagates across a fixed distance in the vasculature. It was approximated based on pulse waveform phase differences between two regions of the face. The propagation speed of pressure pulses is largely determined by arterial stiffness, and arterial stiffness is associated with blood pressure [28]. **Pulse rate variability** provides information about the state of the autonomic nervous system (sympathetic-parasympathetic balance), and greater sympathetic tone is associated with blood pressure increases [29]. Finally, **pulse rate** increases typically co-occur with blood pressure increases since both are key mechanisms for increasing cardiac output to meet the demands of the body. Thus, pulse rate contains information about blood pressure to the degree that it is signaling increases in cardiac output.

Non-blood flow ("meta") features included physical characteristics like gender, age, weight, height, race, and skin tone (as per the six-point

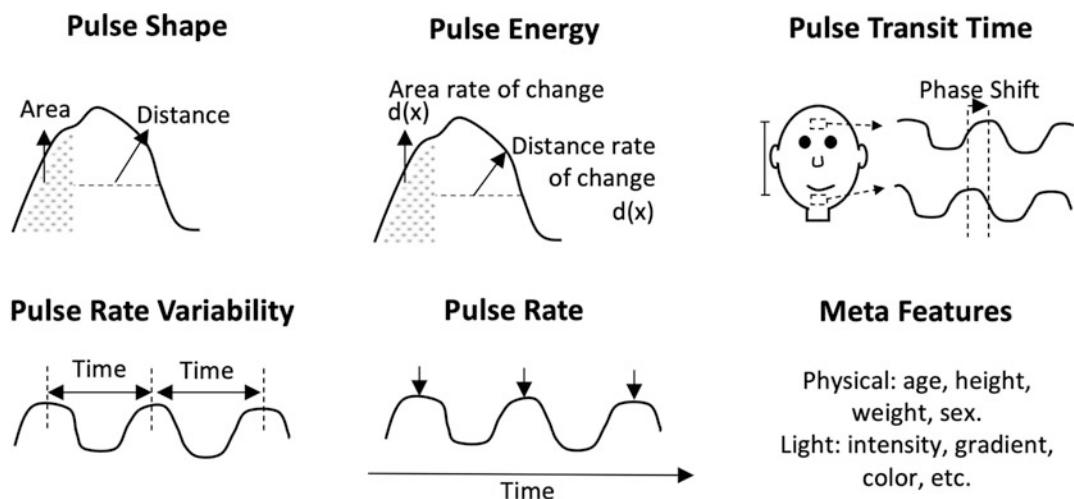


Fig. 3 Feature types used in predicting blood pressure

Fitzpatrick scale), as well as features to help normalize for different lighting conditions (e.g., color information and lighting gradient information).

Predicting Blood Pressure Through the Efficient Combination of Feature Information Using Artificial Intelligence

The Luo [14] and Yang [17] studies extracted features from TOI signal and questionnaires and then input them into a computational model trained to predict either systolic or diastolic brachial blood pressures using advanced machine learning techniques. These models were trained to predict absolute brachial (upper arm) blood pressures and did not require prior calibration with a cuff-based device. Advanced machine learning techniques have thus been used as a method of efficiently combining information from multiple features to accurately predict blood pressure.

Present Advances Using Transdermal Optical Imaging

Accurate Blood Pressure Measurement

In their proof-of-concept study published in *Circulation: Cardiovascular Imaging*, Luo and colleagues demonstrated that combining the predictive power of multiple features makes it possible to accurately estimate blood pressure

[14]. Their blood pressure prediction models trained using advanced machine learning techniques predicted systolic and diastolic blood pressures in a multiracial sample of subjects under controlled conditions when trained and tested on subjects within the normotensive systolic blood pressure range (those with a systolic blood pressure of 100–139 mmHg). Their models predicted with 94.81% accuracy for systolic blood pressure and 95.71% accuracy for diastolic blood pressure. The average measurement error and its standard deviation were 0.39 ± 7.30 mmHg for systolic pressure and -0.20 ± 6.00 mmHg for diastolic pressure, which fall within a generally accepted measurement device limit of 5 ± 8 mmHg [30]. The feature novelty, feature diversity, and large dataset used in this study may have contributed to the high prediction accuracy in this study relative to other remote photoplethysmography approaches to measuring blood pressure.

This system was subsequently incorporated into a smartphone software application called Anura™ (Nuralogix Corporation) that adds additional controls to ensure good measurements under real-world conditions. Specifically, these controls further help account for motion, variable lighting conditions, and skin tone variations through camera calibration (exposure, white balance), measurement constraints (for motion, lighting, and signal quality), and signal normalization techniques [17]. This system is highly robust in

accounting for real-world conditions and informing the user if measurement conditions are inadequate.

In a subsequent study, Yang and colleagues set out to expand the blood pressure prediction range of the models. To do so, they collected new data that included subjects with hypotensive and hypertensive blood pressures and then retrained the blood pressure prediction models on all the data (model performance in a given blood pressure range is largely a function of the quantity of training data in that range). These new models were trained against reference blood pressures measured by a trained observer using the auscultatory technique. As such, they were able to validate these new models in close conformity with internationally accepted blood pressure device validation guidelines published by the Association for the Advancement of Medical Instrumentation (AAMI) [30]. A published study of their preliminary work found that accuracy surpassed AAMI accuracy criteria for both systolic and diastolic blood pressures [17], with an average measurement error \pm error standard deviation of -0.4 ± 6.7 mmHg for systolic blood pressure and 1.2 ± 7.0 mmHg for diastolic blood pressure (both were below the 5 ± 8 mmHg limit). This suggests that Anura is a viable blood pressure measurement tool for a range of blood pressures. A limitation of this work was that it was conducted entirely on Chinese patients, and it is yet unclear how this performance will translate to other races. The process of Anura blood pressure prediction is described in Fig. 4.

Accurate Heart Rate and Heart Rate Variability Measurement

Wei and colleagues (2018) validated heart rate and heart rate variability measured by TOI (as pulse rate and pulse rate variability) against heart rate and heart rate variability measured by the clinical gold standard ECG [13]. They quantified heart rate as pulses per minute and heart rate variability with the SD1/SD2 ratio. SD1 tracks “short-term” variability and SD2 tracks “long-term” variability; higher values of SD1/SD2 imply increased sympathetic activity [31]. Pulse rate measured by TOI was perfectly correlated with heart rate measured by ECG (Pearson $r = 1.0$) [13]. Pulse

rate variability measured by TOI was highly correlated with heart rate variability measured by ECG (Pearson $r = 0.89$). TOI technology therefore captures all heartbeats and accurately captures the subtle variation in timing between them.

Sympathetic nervous system activity as reflected by increases in heart rate variability measures like SD1/SD2 ratio is associated with mental stress [32], and so it is apparent that TOI can also successfully track mental stress. The study results further imply that TOI can also capture gross rhythm abnormalities, although this has not yet been directly investigated.

These TOI-based measures have already been implemented as software-based tools. For instance, the Anura™ smartphone app (Nuralogix Corporation) measures heart rate, heart rate variability (SDNN), and mental stress (a proprietary stress score from 1 to 5 based on heart rate variability) and counts the number of irregularly spaced heartbeats (e.g., heartbeats suggestive of premature ventricular contractions) during its 30-s measurement.

Trends in Medicine and the Potential Impact of TOI

Growing Challenges for Healthcare Delivery

Population aging is expected to bring a rise in the prevalence of age-related chronic diseases (e.g., cardiovascular diseases), and this in turn will increase the need for comprehensive screening, diagnostic and monitoring approaches [33]. Such approaches help mitigate the health burden of disease, but they are highly resource-intensive when they are carried out in the clinic or long-term care home setting where they require time from primary care physicians and other staff. Growing primary care physician shortages in many jurisdictions will limit the ability to deliver this type of care. Concurrently, patients are demanding a more “patient-centered” approach to their care that is more convenient and gives them more control and involvement in their care. Issues with accessibility might further limit the ability to deliver high-quality and efficient care, particularly when it comes to remote communities

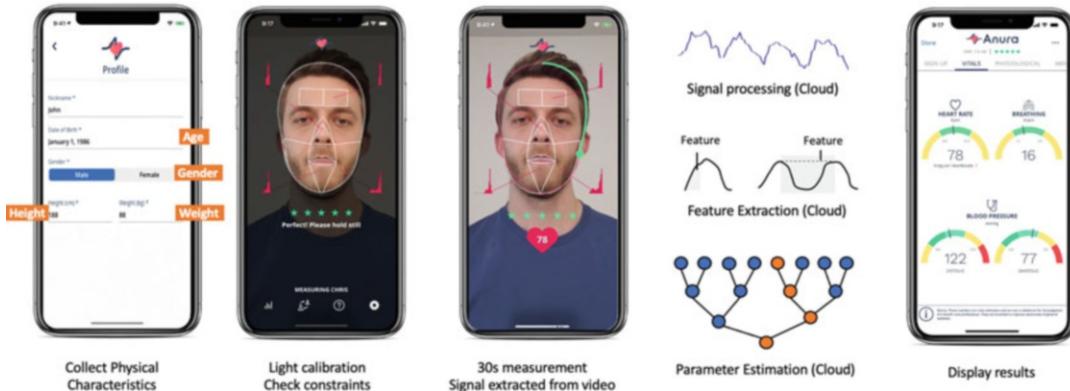


Fig. 4 Process of transdermal optical imaging in prediction of blood pressure with the Anura™ smartphone app (Nuralogix Corporation). Physical characteristics are collected on the profile screen during sign-up. Entering measurement mode triggers exposure and white balance calibration. Distance, motion, and lighting constraints are checked to ensure that conditions are adequate before beginning the measurement. A star-based signal quality system rates the lighting conditions (five stars indicate

and providing care safely during times of elevated infectious disease risk (e.g., when restrictions are in place for COVID-19).

Improving Healthcare Quality and Efficiency with Patient-Centered Health Monitoring

A shift from health monitoring in the clinic to self-monitoring at home would facilitate comprehensive monitoring of cardiovascular parameters like blood pressure and heart rate at reduced cost. Comprehensive monitoring maintains health and reduces costs associated with illness [34]. For instance, facilitating hypertension awareness and control greatly reduces the risk of cardiovascular diseases. Further, remote health monitoring in heart failure patients is credited with reducing costly hospital readmissions in heart failure patients by more than half through identifying complications early [35]. A shift away from clinic-based monitoring would improve convenience (via fewer trips to the clinic) and allow many with chronic diseases to remain in their own homes and in the community for longer before requiring care at a long-term care home [36]. Payers are realizing massive cost savings with such models and are increasingly

ideal conditions and will result in the most accurate measurements). The measurement will then begin and proceed for 30s. Blood flow signal is extracted from video during this time. Signal is then uploaded to the cloud for further processing and feature extraction. Machine learning-based models in the cloud estimate parameters. Finally, these parameters are downloaded to the phone and displayed to the user

incentivizing them in favor of costly long-term care arrangements [37, 38]. In fact, it is estimated that remote health monitoring will reduce chronic disease management costs by 10–20% or more. Finally, patients have shown a willingness to engage in patient-centered remote monitoring solutions for both health and wellness. Technological innovations in remote home monitoring will continue to be a major enabler for self-monitoring at home [33].

Improving Healthcare Accessibility with Telemedicine

Telemedicine is beginning to address issues of accessibility and convenience in medicine by assessing, diagnosing, monitoring, and communicating with patients remotely, most typically through video conferencing technology. Its use has been growing as a way to increase convenience (through fewer trips to the doctor) and provide high-quality care to remote communities. Most recently during the COVID-19 pandemic it has been used as a substitute for in-person visits to mitigate the risk of spreading infection. While the proliferation of ubiquitous high-speed internet and computing devices has enabled this technology, its application has thus far been limited by the

lack of access to standard cardiovascular parameters (e.g., blood pressure, heart rhythm) that are commonplace in the clinic but most often cannot be measured at home. For telemedicine to reach its full potential, it will be necessary for measurement tools to match the ubiquity of video conferencing tools.

TOI-Based Tools Could Transform Personalized Self-Monitoring and Telemedicine

TOI technology has unique characteristics that would enable the creation of tools that address obstacles to measuring cardiovascular parameters at the patient level, as well as address wider challenges in healthcare delivery. Contactless measurement of blood pressure, heart rate, heart rhythm, heart rate variability, and stress via TOI would be much more comfortable than inflatable cuffs, ECG leads, or even wearables, for instance. Since TOI acquires pressure pulse information continuously, it can average out slowly oscillating physiological waves (e.g., Mayer waves) and determine blood pressure not only more precisely but in one quick measurement instead of the three long measurements required with cuff-based devices. Importantly, TOI technology could be implemented on any device capable of capturing and processing video. Such devices are already ubiquitous (e.g., smartphones, tablets, computers). Since the hardware required to run TOI is already everywhere (e.g., at home, at work, and on a video call with a doctor), measuring cardiovascular characteristics using TOI is highly convenient and does not require purchasing or carrying around extra equipment like wearables or often bulky cuff-based devices.

These improvements in comfort, convenience, and access to measurements have the potential to address challenges in healthcare delivery by facilitating patient-centered health monitoring and telemedicine (Fig. 3). TOI-based tools could facilitate a patient-centered approach and improve healthcare quality by making physiological measurements more accessible, more convenient, and more resource efficient. For instance, in the general public unaware of their blood pressure status,

unprecedented access to clinical-grade blood pressure measurement right on one's mobile phone could facilitate the detection of abnormal blood pressure and help drive much-needed increases in hypertension awareness. Diagnosed hypertensives could benefit from increased comfort and convenience over traditional cuff-based devices in monitoring response to therapy. Facilitating regular measurements any time and any place (e.g., even outside the home) could provide greater awareness of high blood pressure triggers throughout the day and help drive increases in hypertension control (via lifestyle changes or medication). This would help address the massive health and financial burden of cardiovascular disease. More comfortable and convenient measurement tools are liable to drive further uptake of remote monitoring programs and thus reduce costs at the healthcare delivery level. TOI technology within telemedicine could enable the measurement of cardiovascular parameters that were previously not possible to measure over video calls unless the patient had specialized equipment at home. Such an advancement would drastically increase the utility of video consultations. A summary of TOI-based functionality is shown in Fig. 5.

Future Uses and Challenges for TOI

Given the tight interrelation between the cardiovascular system and other physiological parameters, plethysmographic signal acquired via TOI can further estimate additional parameters to varying degrees of accuracy. For instance, the expansion and contraction of the diaphragm during respiration exerts significant pressure modulations on the cardiovascular system and enables the quantification of respiration rate. The Anura smartphone app (Nuralogix Corporation) estimates respiration rate from plethysmographic signal, as do some clinical-grade pulse oximetry tools. Other health parameters convey information through the vascular system and might also be measured. They include blood oxygen saturation (by quantifying differences in the absorption spectra of oxygenated versus deoxygenated

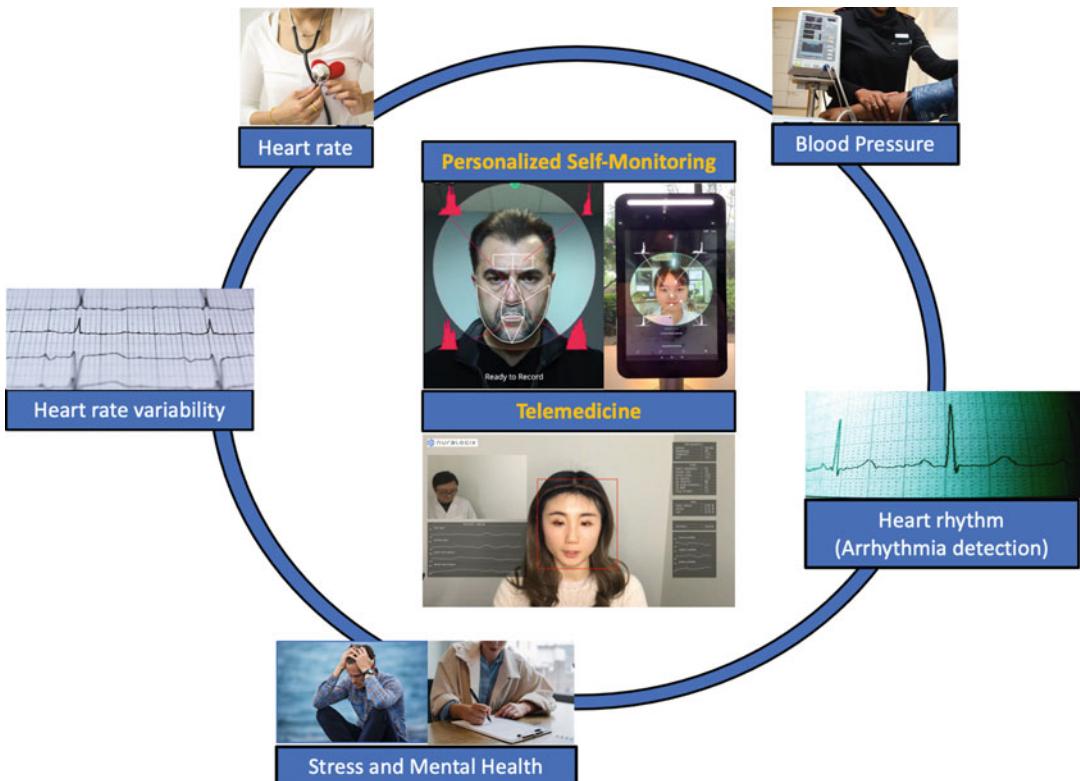


Fig. 5 TOI-based functionality in personalized self-monitoring and telemedicine applications. TOI has enabled the quick, comfortable, and convenient measurement of

parameters like heart rate, heart rate variability, stress, blood pressure, and heart rhythm using already ubiquitous hardware like the smartphone

hemoglobin), blood biomarkers (e.g., hemoglobin concentration, cholesterol level, and even hemoglobin A1c [39]), disease risks (e.g., hypertension, diabetes mellitus, obesity), and even mental health conditions (e.g., depression and anxiety via the effects of autonomic activity on blood flow).

Employing TOI-based measurement tools in the clinic will require medical device approval from appropriate regulatory bodies (e.g., the Food and Drug Administration in the United States). These agencies commonly cite specific validation standards for blood pressure and heart rate monitoring equipment. However, such standards are technology-specific and currently there are no specific standards for remote photoplethysmography devices.

The most closely related blood pressure device standard from the Association for the Advancement of Medical Instrumentation (AAMI) [30]

provides guidelines for validating cuff-based automated blood pressure devices against a gold standard reference device (e.g., auscultation by a trained observer). Despite technological differences, many features of this standard are likely to be translatable to TOI, including testing conditions, subject distribution requirements (for sex, age, and blood pressures), validation procedure, and accuracy requirements. A major accuracy criterion is that the mean error and error standard deviation between the systolic and diastolic blood pressures of reference and test devices fall below 5 mmHg and 8 mmHg, respectively. The technological differences between image-based devices like TOI and cuff-based devices may necessitate consideration of additional factors during validation like skin tone, lighting conditions, and motion, as such factors could conceivably impact measurements in an image-based system. Regulators have also recognized validation standards for

heart rate measurement devices. For instance, the accuracy requirement for heart rate measured by a pulse oximeter is an error of less than 10% or ± 5 beats per minute (whichever is greater) [40]. Once again, there are no guidelines specific for remote video-based technology like TOI, and so manufacturers will need to seek specific guidance from regulators on the validation approach to be used.

Finally, given the pivotal role of blood pressure and heart rate measurement in various clinical applications, it will be crucial to assess how TOI-based tools compare with existing standard techniques in terms of practical effectiveness. For instance, pragmatic trials might investigate how well measurements agree with existing standard methods and the impact of this on diagnostic or treatment decisions. With the necessary regulatory approvals and clinical trials, TOI-based tools will be able to provide accurate, convenient, and contactless screening, diagnosis, and monitoring of patient health. As a result, these tools will have the potential to revolutionize healthcare delivery by enhancing access and efficiency, which in turn will not only reduce healthcare costs but also improve the health of people all over the world.

References

- Mills KT, Bundy JD, Kelly TN, Reed JE, Kearney PM, Reynolds K, et al. Global disparities of hypertension prevalence and control. *Circulation*. 2016;134(6):441–50.
- Barszczuk A, Yang D, Wei J, Huang W, Feng Z-P, Lee K, et al. Potential impact of the 2017 high blood pressure guideline beyond the United States: a case study of the People's Republic of China. *Am J Hypertens*. 2020;33(9):846–51.
- Oparil S. Global blood pressure screening: a wakeup call. *Hypertension*. Lippincott Williams and Wilkins. 2020;76:318–20.
- Whelton PK, Carey RM, Aronow WS, Casey DE, Collins KJ, Dennison Himmelfarb C, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Pr. *J Am Coll Cardiol*. 2018;71(19):e127–248.
- Mahdi A, Watkinson P, McManus RJ, Tarassenko L. Circadian blood pressure variations computed from 1.7 million measurements in an acute hospital setting. *Am J Hypertens*. 2019;32(12):1154–61.
- Bartels K, Esper SA, Thiele RH. Blood pressure monitoring for the anesthesiologist. *Anesth Analg*. 2016;122(6):1866–79.
- Ponikowski P, Spoletini I, Coats AJ, Piepoli MF, Rosano GM. Heart rate and blood pressure monitoring in heart failure. *European Heart Journal Supplements*. 2019;21(Supplement_M):M13–16.
- Stauss HM. Heart rate variability. *Am J Physiol – Regul Integr Comp Physiol*. 2003;285(5):R927–R931.
- Ahmad S, Tejuja A, Newman KD, Zarychanski R, Seely AJE. Clinical review: a review and analysis of heart rate variability and the diagnosis and prognosis of infection. *Crit Care*. 2009;13(6):1–7.
- Kleiger RE, Thomas Bigger J, Bosner MS, Chung MK, Cook JR, Rolnitzky LM, et al. Stability over time of variables measuring heart rate variability in normal subjects. *Am J Cardiol*. 1991;68:626.
- Kim HG, Cheon EJ, Bai DS, Lee YH, Koo BH. Stress and heart rate variability: a meta-analysis and review of the literature. *Psychiatry Investig*. Korean Neuropsychiatric Association. 2018;15:235–45.
- Liu J, Luo H, Zheng PP, Wu SJ, Lee K. Transdermal optical imaging revealed different spatiotemporal patterns of facial cardiovascular activities. *Sci Rep*. 2018;8(1):10588.
- Wei J, Luo H, Wu SJ, Zheng PP, Fu G, Lee K. Transdermal optical imaging reveal basal stress via heart rate variability analysis: a novel methodology comparable to electrocardiography. *Front Psychol*. 2018;9:98.
- Luo H, Yang D, Barszczuk A, Vempala N, Wei J, Wu SJ, et al. Smartphone-based blood pressure measurement using transdermal optical imaging technology. *Circ Cardiovasc Imaging*. 2019;12(8):e008857.
- Barszczuk A, Lee K. Measuring blood pressure: from cuff to smartphone. *Curr Hypertens Rep*. 2019;21(11):1–4.
- Mukkamala R. Blood pressure with a click of a camera? *Circulation: Cardiovasc Imaging*. 2019;12(8): e009531–e009531.
- Yang D, Xiao G, Wei J, Luo H. Preliminary assessment of video-based blood pressure measurement according to ANSI/AAMI/ISO1060-2:2013 guideline accuracy criteria: Anura smartphone app with transdermal optimal imaging technology. *Blood Press Monit*. 2020;25(5):295–298.
- Gallagher D, Adji A, O'Rourke MF. Validation of the transfer function technique for generating central from peripheral upper limb pressure waveform. *Am J Hypertens*. 2004;17(11):1059–1067.
- Verkruyse W, Svaasand LO, Nelson JS. Remote plethysmographic imaging using ambient light. *Opt Express*. 2008;16(26):21434–21445.

20. Kamshilin AA, Margaryants NB. Origin of photoplethysmographic waveform at green light. *Phys Procedia*. 2017;86:72–80.
21. Takano C, Ohta Y. Heart rate measurement based on a time-lapse image. *Med Eng Phys*. 2007;29(8):853–857.
22. Lewandowska M, Rumiński J, Kocejko T, Nowak J. Measuring pulse rate with a webcam – a non-contact method for evaluating cardiac activity. In: 2011 Federated conference on computer science and information systems, FedCSIS 2011. 2011.
23. Jain M, Deb S, Subramanyam AV. Face video based touchless blood pressure and heart rate estimation. In: 2016 IEEE 18th international workshop on multimedia signal processing, MMSP 2016. 2017.
24. De Haan G, Jeanne V. Robust pulse rate from chrominance-based rPPG. *IEEE Trans Biomed Eng*. 2013;60(10):2878–2886.
25. Moço A V, Stuilk S, de Haan G. Motion robust PPG-imaging through color channel mapping. *Biomed Opt Express*. 2016;7(5):1737–1754.
26. Adachi Y, Edo Y, Ogawa R, Tomizawa R, Iwai Y, Okumura T. Noncontact blood pressure monitoring technology using facial photoplethysmograms. In: Proceedings of the annual international conference of the IEEE engineering in medicine and biology society, EMBS. 2019.
27. Drummond PD. Psychophysiology of the blush. In: The psychological significance of the blush. Cambridge University Press; 2009.
28. Elgendi M. On the analysis of fingertip photoplethysmogram signals. *Curr Cardiol Rev*. 2012;8(1):14–25.
29. Fisher JP, Paton JFR. The sympathetic nervous system and blood pressure in humans: implications for hypertension. *J Hum Hypertens*. Nature Publishing Group. 2012;26:463–75.
30. Association for the Advancement of Medical Instrumentation. ANSI/AAMI/ISO 81060-1:2007/(R)
- 2013 Non-invasive sphygmomanometers – Part2: Clinical investigation of automated measurement type. 2013.
31. Roy B, Ghatak S. Nonlinear methods to assess changes in heart rate variability in Type 2 diabetic patients. *Arq Bras Cardiol*. 2013;101:317–327.
32. Melillo P, Bracale M, Pecchia L. Nonlinear heart rate variability features for real-life stress detection. Case study: students under stress due to university examination. *Biomed Eng Online*. 2011;10:96.
33. GSMA MC. mHealth: a new vision for healthcare [Internet]. 2012. <https://www.gsma.com/iot/wp-content/uploads/2012/03/gsmamckinseyhealthreport.pdf>
34. Tarride J-E, Lim M, DesMeules M, Luo W, Burke N, O'Reilly D, et al. A review of the cost of cardiovascular disease. *Can J Cardiol*. 2009;25(6):e195–202.
35. Ottawa Heart Institute: Telehome Monitoring [Internet]. <https://www.ottawaheart.ca/healthcare-professionals/regional-national-programs/telehome-monitoring>
36. Broderick A, Lindeman D. Scaling telehealth programs: lessons from early adopters. *The Commonwealth Fund*. 2013;1654(1):1–10.
37. Lakhdar K, Black G. Blurring the lines: convergence in Canadian Health & Life Sciences [Internet]. KPMG. <https://assets.kpmg/content/dam/kpmg/pdf/2016/05/BlurringtheLines-Convergence-in-Canadian-HLS.pdf>
38. Kayali B, Kimmel Z, van Kuiken S. Spurring the market for high-tech home health care. McKinsey & Company [Internet]. <http://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/spurring-the-market-for-high-tech-home-health-care>
39. Lyons TJ, Basu A. Biomarkers in diabetes: hemoglobin A1c, vascular and tissue markers. *Transl Res*. 2012;159(4):303–312.
40. Association for the Advancement of Medical Instrumentation. ANSI/AAMI/EC13:2002 Cardiac monitors, heart rate meters, and alarms. 2002.



Dina Radenkovic, Alex Zhavoronkov, and Evelyne Bischof

Contents

Introduction	1158
The Advent of Deep Aging Clocks	1159
Federated Learning for Biomarker Discovery and Development	1160
Longevity Physicians: Emerging Specialists and the Need of a Tailored Education	1161
Healthy Versus Wealthy Longevity: Longevity Medicine and Public Health	1164
AI Applications in Medicine: The Fundament of Precision Medicine	1165
Conclusion and Future Perspectives	1166
References	1166

D. Radenkovic

Hooke London, London, UK

King's College London, London, UK

Buck Institute for Research on Aging, Novato, CA, USA

e-mail: drdina@hooke.london

A. Zhavoronkov (✉)

Insilico Medicine, Hong Kong Science and Technology

Park, Hong Kong, China

Three Exchange Square, The Landmark, Deep Longevity,
Inc, Hong Kong, China

Buck Institute for Research on Aging, Novato, CA, USA

e-mail: alex@insilico.com

E. Bischof

International Center for Multimorbidity and Complexity in
Medicine (ICMC), Universität Zürich, Schweiz, Zurich,
SwitzerlandCollege of Clinical Medicine, Shanghai University of
Medicine and Health Sciences, Shanghai, China

Human Longevity, Inc., San Diego, CA, USA

Abstract

Since 2013, deep learning systems, a form of artificial intelligence (AI), outperformed humans in image, voice, and text recognition, video games, and many other tasks. In medicine, AI has outperformed humans in dermatology, ophthalmology, and several areas of diagnostic medicine. Since the first publication of aging clocks based on methylation data in 2013 by Horvath and Hannum, AI techniques were used to predict human age, mortality, and health status using blood biochemistry in 2016 and later using transcriptomics, proteomics, imaging, microbiome, methylation, activity, and even psychological survey data. Today, these deep aging clocks (DACs) are being used by the research physicians to evaluate the effectiveness of longevity interventions, clinical trial enrollment and monitoring, risk profiling, biological target identification, and personalized medicine. The advent of data type-specific and multi-omics DACs allowed for the nascent field of aging clock-driven preventive and regenerative medicine, referred to as longevity medicine, to emerge.

Introduction

Over the past decade, we witnessed unprecedented advances in the field of biogerontology and the massive convergence of biotechnology, information technology, AI, and medicine. The birth of longevity medicine, which integrates the latest advances in many of these fields of science and technology, is not surprising but rather embraced by progressive clinicians, scientists, and patients. Longevity medicine is advanced personalized preventative medicine powered by deep biomarkers of aging. This domain is extremely novel – aging clocks were first published in 2013 by Steven Horvath et al. [1] and deep aging clocks first published in 2016 by Alex Zhavoronkov et al. [2]. Nevertheless, it became one of the most important areas of

precision medicine. What started with a symbiotic effort of mathematics and biochemistry evolved to an artificial neural network approach – deep learning. The method is used to develop deep age predictors with the potential to accelerate research and clinical translation of causal relationships in nonlinear systems. Applying the clocks in the clinic may allow clinicians to determine the efficacy and efficiency of interventions and prognostic and preventative measures. A tailored longevity medicine education is incumbent in order to train clinical experts in longevity medicine, who will play a crucial role, since the universal process of aging renders every human a patient of longevity medicine, therefore preventing the multimorbidity characterized by old age, including physical and/or mental limitations. Drug therapy poses a further problem of multimorbidity. This is because older people often take many different medicines, and this can also increase the number of undesirable side effects (Fig. 1).

Longevity medicine is suited for advances in machine learning for numerous reasons. Longevity is affected by a large number of intrinsic factors such as genetics and epigenetics [3] and by extrinsic factors such as medical history, diet, exercise, socioeconomic determinants of health, and geolocation [4, 5]. Additionally, the effects of these exposures summate over the course of a lifetime affecting, among other factors, development or progression of age-related disease. Collating this data on biomarkers and drug candidates using a trial-by-trial basis conducted on sufficiently large patient or participant cohorts over long-term follow-up is resource-intensive and unpragmatic. For such heterogeneous and complex data, machine learning is a well-suited candidate both to identify suitable biomarkers in advance of physical trials and to do so with higher accuracy than traditional statistical analysis.

A quintessential tense of a machine learning algorithm is endpoint selection. Actual age and clinical outcomes such as the development or progression of age-related disease provide a

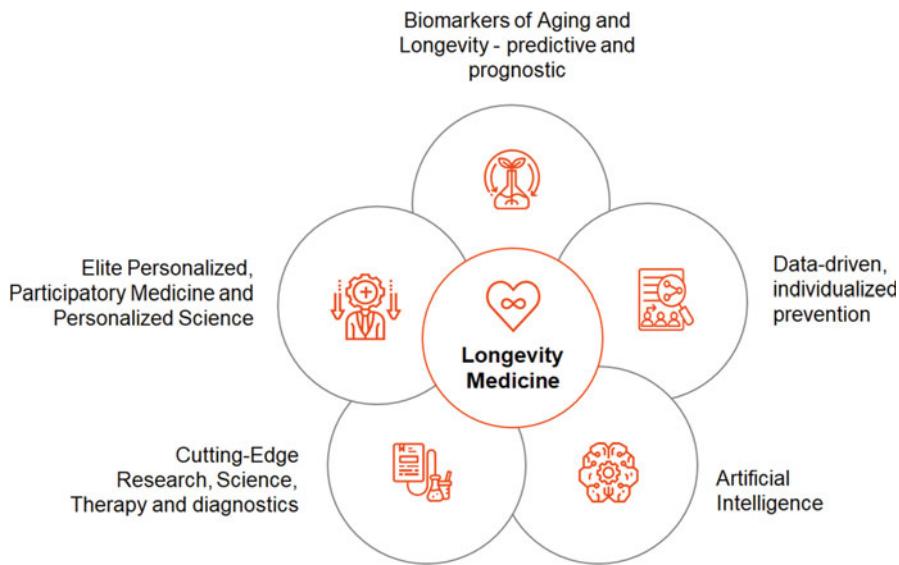


Fig. 1 The technologies enabling the field of longevity medicine

limited view of one's age status. Biological age, which more closely explains age status in the context of one's physiology, is a measure that focuses on progression through aging instead of actual age. Surrogate biomarkers of biological age, termed deep aging clocks, have been shown to estimate disease mortality [6], discern the age of developmental tissue [7], and estimate time remaining before onset of age-related pathologies [8]. The utility of these tools is often based on multi-omics or panomics databases with extensive measures generated on trillions of data points. The interpretation and analysis of these datasets requires advanced techniques for which machine learning may be better suited over traditional methods.

It is often difficult to train an accurate model based on a single dataset as this makes the model prone to bias and it may not have the desired complexity given the project aims. Additionally, databases are generated interinstitutionally with different protocols and specifications across potentially diverse biological features, which makes some of them prone to having gaps in data provision, necessitating the imputation of missing or unknown values. Access to multiple

datasets will negate these effects, but this is practically near impossible due to inherent issues of confidentiality and governance surrounding sensitive medical data. Therefore, techniques like federated learning [9] that train unified models across multiple databases without sharing the data sources would overcome some of the strong barriers around patient confidentiality. This may also bridge together wisdom gained from the academic and medical fields with wisdom generated from data collected in the private sector, which is traditionally difficult due to large conflicts of interest.

The Advent of Deep Aging Clocks

Many biological characters have demonstrated broad correlations with chronological age, including telomere attrition and racemization of amino acids in proteins [10]. The primary progenitor, however, in the advent of aging clocks is what would become known as the “epigenetic clock.” In 2013, Steve Horvath et al. established significant correlations between CpG methyl groups and chronological age, suggesting that this deep aging

clock demonstrated the cumulative effect of an epigenetic maintenance system [1]. This has led to significantly increased attention in the fields of longevity, biogerontology, biomarkers of aging, and the role of machine learning in consolidating these data into precise and coherent models.

What was not well established were geroprotective interventions that could reduce the rate of aging in the novel epigenetic clock and whether these could translate to the actual lengthening of both healthspan and lifespans. Hence, research proliferated utilizing hematological biomarkers indicative of morbidity and mortality putatively involved with aging. This would include a hematological, biochemical clock developed by Putin et al. [2]. Through the use of deep neural networks (DNN), the group were able to train AI to produce a robust model utilizing 41 factors that could predict one's chronological age with impressive accuracy ($R^2 = 0.82$, MAE = 5.5). The benefit of employing DNN architecture in model generation in this instance was that it was able to recursively remodel and account for non-linear associations among features. Importantly, they established for their model the five most important markers for predicting human chronological age: albumin, glucose, alkaline phosphatase, urea, and erythrocytes. What was particularly useful for this deep aging clock was that interventions that could ameliorate the biomarkers used in the model were already well established, and hence, some of the first truly geroprotective interventions were identified.

Indeed, the advent of deep biomarkers of aging has led researchers to target different loci in human and animal biology. For example, researchers as early as 2015 were investigating the role of gene expression through relative quantitative changes in mRNA presence for particular genes and correspondent proteins [11]. Using a cohort of 14,983 individuals, they identified 1497 genes that are differentially expressed with chronological age and used this to develop a clock with a 7.8-year MAE. Importantly, they established that individuals for whom transcriptional differentiations departed above the model's slope exhibited biological features associated with aging, notably, blood pressure, cholesterol levels,

fasting glucose, and body mass index. They also established that their gene panel was enriched for the presence of potentially functional CpG-methylation sites in enhancer and insulator regions that associate with both chronological age and gene expression levels, therefore discovering a mechanism by which the epigenetic clock (DNAm age) could behave as an actor in the development of morbidities with age.

Another transcriptome aging clock was produced in 2018 by Mamoshina et al. [12]. In this instance, they employed the use of DNNs to predict chronological age from 545 gene expression profiles in skeletal muscle of healthy individuals. Among currently published transcriptomic clocks, this clock is the most accurate, demonstrating an MAE of 6.24 years, which may indicate that transcriptomic age prediction requires more complex machine learning techniques than those commonly used in DNAm clocks [13].

Federated Learning for Biomarker Discovery and Development

One's longevity is dependent on a large number of covariates such as dietary habits, exercise levels, and socioeconomic determinants of health. As a result, unbiased long-term trials are used to identify or repurpose new or existing agents that contribute toward greater longevity. For example, the Targeting Aging with Metformin (TAME) trial [14] investigates the use of metformin to delay development or progression of age-related chronic disease in 3000 participants aged between 65 and 79 for 6 years. Aging is chronic and develops in conjunction with long-term environmental exposures. Thus, it may be a further 10 or more years posttrial for the clinical endpoints to appear in participants before the impact toward longevity can be fully evaluated.

The long time-to-event in trials may come at high cost for longevity drug discovery. Machine learning (ML) algorithms may be employed to generate advanced accurate predictions into the efficacy of interventions, particularly for repurposing existing medicines. Data on existing medicines is collected in electronic health records

and the generation of research repositories. This applies especially for commonly prescribed medicines such as metformin.

ML is well suited to the complex variability of human studies and is already being applied for fields such as drug target discovery and protein folding, but a limitation into their application is often a shortage of usable data [15]. Large panomics databases are generated to provide ample numbers of data points for medical research, such as the UK BioBank [16] comprising 500,000 participants. Despite the large number of participants, when variables of interest are selected such as a given disease and when covariates such as ethnicity are stratified, the usable numbers of participants may be insufficient to train an accurate machine learning algorithm. Furthermore, curated datasets may also have biases stemmed from processes such as participant acquisition. To address these limitations, it may be tempting to consolidate different datasets in order to have a sufficiently sized dataset including all variables of interest generated from a variety of sources to counter bias that would be present in a single dataset.

Medical data is highly sensitive and there are many regulations over how it is used [17]. Anonymization by removing personally identifiable information is generally insufficient to surpass this barrier making it practically difficult to combine medical datasets. Another reason why medical data is unstandardized is that generating shared datasets is time intensive and may be costly.

Federated learning [9] is a ML algorithmic methodology which was created for collaborative model generation where the learned wisdom is shared but the data is not shared between repositories. This means that a single model can be generated without breaking data governance over a large number of datasets. Federated learning by design consists of using a central server to distribute nodes for each database source behind the owner institution's firewall. The local nodes train the model and return it to the central server for aggregation and redistribution to the local nodes. Split learning, a type of federated learning, uses a peer to peer design, where there is no

central server and the individual nodes share their trained models, aggregate them locally, and redistribute them with each other. These protocols for both designs are repeated until the training is completed [18]. Federated learning performs faster than split learning as the clients generate models in parallel, but split learning has improved privacy as the architecture is split between the clients and the server but is more computationally intensive on each local node [19].

In addition to federated learning, other methods can be employed either as alternatives when federated learning is unavailable or to help validate a federated learning model. The bias and variability of single nonfederated models trained on small datasets can be compensated against using traditional statistical methods. ML models on small data can have a tendency to overfit training data and may not interpret unknown data accurately, but using models such as logistic regression on expert selected features and selectively removing outliers may improve the efficacy of the model. This methodology may be further optimized by running multiple models and generating a weighted average [20].

Longevity Physicians: Emerging Specialists and the Need of a Tailored Education

With the progress in geroscience and biogerontology, as well as in AI-based diagnostic and therapeutic tools, longevity medicine as a field has been increasingly in demand by educated patients. For longevity medicine to become an organic expertise area or, preferably, a specialty, just as oncology or cardiology, it needs to be practiced. This in turn implies a solid education of physicians, allowing them to not only acquire the necessary fundamentals, such as hallmarks of aging, but also to rapidly internalize and apply related research outcomes and tools.

Furthermore, as to develop and adequately practice longevity medicine, an ongoing continuous learning is required, encompassing the likewise rapidly evolving areas of precision, prevention, and functional medicine. Ultimately,

educating physicians in longevity medicine is one of the most challenging endeavors, requiring a structured, meticulous approach of interdisciplinary educators. Overall, physicians' continued education is still inadequate to meet the challenges and opportunities of longevity medicine.

Current medical training does not include AI, not to mention machine or deep learning [21]. The AI-enforced tools for early diagnostics and prevention of communicable and non-communicable (NCD) diseases are thus unpopular among the majority of physicians [2, 22, 23]. The so-called millennials are increasingly exposed to innovative solutions in medicine and to active patients' inquiries. Facing a lack of structured, physician-oriented, conceptualized curricula, this new generation, with a tremendous potential for global healthy longevity, is deserting and often diffusing away from the field. At the same time, a growing interest in longevity medicine is observed among related fields including biotechnologists, biologists, geroscientists, as well as among the general public, pharmacologic companies, and target industry groups. Interestingly, the latter two demonstrate an extremely high interest in growing a longevity physician community, while specific approaches to incentivize specific educational approaches (building of certified courses, curricula, accreditations) require significant resources. As long as academia will not sufficiently support such endeavors, the educational initiatives remain grandly scattered, in a form of individual or lecture series given by mostly nonclinical experts. In addition, despite the benefits longevity-focused medicine has to offer, currently, only its fragmentary aspects are available within the certification system: lifestyle, holistic, integrative medicine, or geriatrics. Extremely time-limited clinicians rarely can accommodate extracurricular educational activities in their schedules. Currently, we mostly face reactive medicine, managing diseases rather than mitigating risks toward pathogenesis. Since most healthcare systems are fragmented, specialists are often dispersed and comprehensive patient care is suboptimal. Most importantly, there is extremely limited access to

the practitioners who have self-initiated their education into longevity sciences, follow the rapid development of various diagnostic and monitoring solutions, and implement these from the longevity standpoint.

The paradigm of longevity and healthy aging as the top priority will greatly impact the primary, secondary, and tertiary prevention rates; it is essential that doctors have the access to well-structured and practitioner-friendly course contents. Development of such courses is only possible through interdisciplinary efforts that would generate contents and contexts teaching both the fundamentals and ways of how they can be implemented in the clinical practice.

Education is the foundation for longevity health findings to be implemented in the daily practice for patients and in preventative settings. It decimates the knowledge gap between physicians and nonclinical longevity experts (bio-/gerontologists, AI and computer scientists, etc.) and reduces stigmatization of longevity medicine being falsely dismissed as a highly anecdotal movement toward life prolongation. In contrast, longevity medicine is a highly scientific approach toward the extension of a healthy and productive lifespan, with an AI-based precision approach using measurable markers of aging.

An accurate physicians' education in longevity medicine must aim to clearly demonstrate current medical practice, which evaluates and optimizes the parameters of patients within the reference range for their corresponding age groups from truly personalized medicine based on large data. Even if an age group is selected based on a variety of further variables, such an approach is at most a personalized one, but not a precise or individual one. Longevity medicine brings together the best practices from various biomedical disciplines and AI to evaluate the patient's biological age throughout his/her course of life in order to reduce the gap between the current and the parameters of maximum physical performance (based on a calculated ideal biological age). A longevity physician is thus required to be able to not only apply measurements of aging clocks but also to then (with AI assistance) identify ways to reduce the gap between the current and the optimal biological age. These ways include, at the moment,

noninvasive lifestyle modifications (e.g., physical activity, intermittent fasting, circadian rhythm readjustment) or geroprotective supplements and anecdotal invasive methods. It is apparent that these approaches will require ongoing data collection in order to customize and improve protocols, as well as individual AI-supported recommendations (Fig. 2).

Only in the year 2020, a global interdisciplinary team developed the first official Longevity Medicine for Physicians course, covering topics of biogerontology, machine learning, biostatistics, differential diagnosis, programming, molecular biology, immunology, geroprotective interventions, drug design, healthcare organization, and others, as well as providing an overview of clinical applications of recent advances in aging research, skills to evaluate the validity of biomarkers of aging and other biological age testing systems, and knowledge of the available longevity therapies to tackle diseases that are mostly based on senescence-related processes in the organisms. It additionally bridges discrepancies in awareness and information on advances in research while bringing practicable examples of implementation into clinical practice. The unprecedented increase in the percentage of people over 65 years of age and corresponding increase in the illness, social, and economic burden associated

with aging require us to advance our understanding of the aging process and how to tackle those processes and provide the care needed. This and similar initiatives will ultimately lead to major benefications in healthcare systems as paramount, effective management of diseases, since progress in longevity and biogerontology research will likely increase the healthy productive lifespan and the number of years of government support in old age. It is therefore incumbent to educate physicians about the most recent parameters that can be applied to established models of prevention of diseases by tackling and applying biogerontology advancements. As such, they will be linked to economic growth via biomedical progress rate, the rate of clinical adoption, and the rate of change in retirement age.

Aging is a complex multifactorial process leading to loss of function, causing multiple NCDs, rendering prone toward CDs and premature mortality [24, 25]. There are many theories explaining the origin of the overall process [26, 27] and cause and effect relationships between different processes and systems, including aging of the immune system [28], inflammation [29–32], fibrosis [33, 34], mineralization of connective tissue [35], cellular senescence [36], wear and tear, and many others. In addition, many genetic and epigenetic changes implicated in aging and

AI-Guided Longevity Medicine

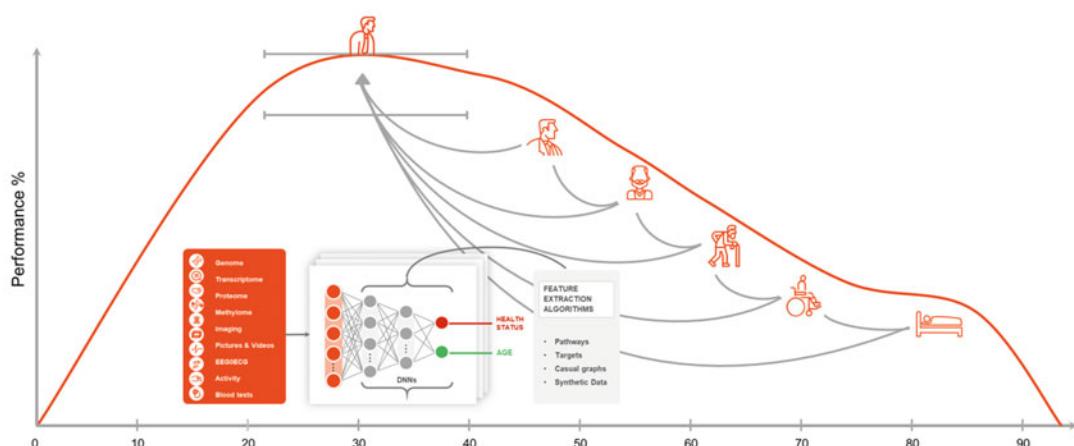


Fig. 2 How human performance changes with age. Panomics technologies integrated with deep neural networks can interpret this and identify geroprotective pathways

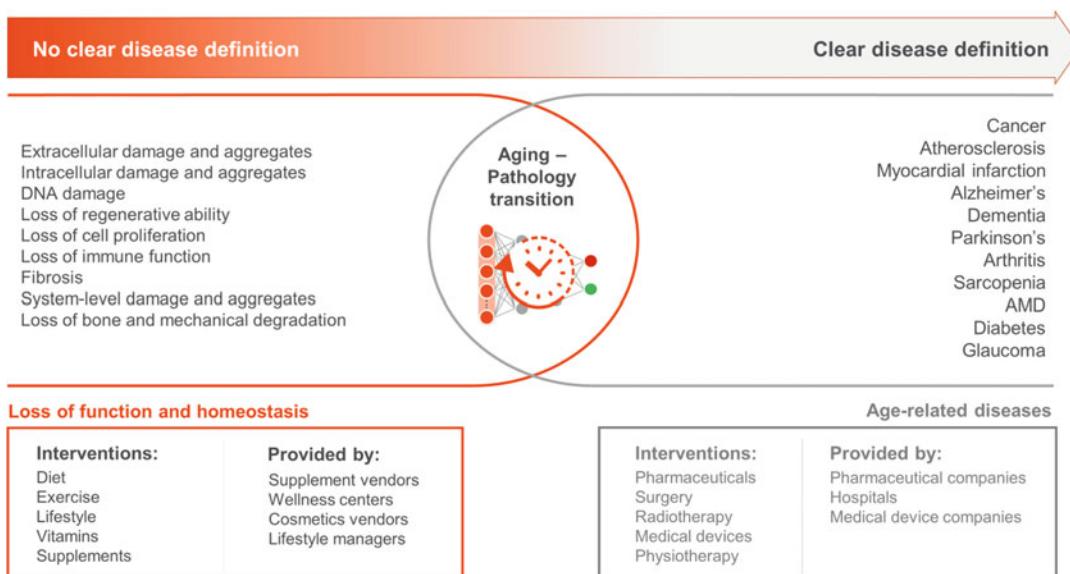


Fig. 3 Cellular processes with no clear disease definition contribute to aging and pathology over time and are affected by environmental factors such as diet. These can then lead to age-related disease

longevity are associated with aging in model organisms [37, 38]. Even though their role and action in human aging are uncertain, many of these changes are also associated with non-communicable diseases [37–43]. Recent discoveries showing that mechanisms involved in cancer are strongly associated with the aging process have led to multiple proposals to prevent cancer and other age-related diseases using drugs that increase lifespan in model organisms [44, 45]. AI and machine learning in the course are an inseparable aspect of modern medicine, especially tackling the prevention of global burden diseases. AI is expected to have a major impact in healthcare, where it can be used for the development of effective personalized medicine based on the interpretation of large medical databases gathered over the years by companies and healthcare providers [22, 46–49] (Fig. 3).

Healthy Versus Wealthy Longevity: Longevity Medicine and Public Health

The debate around longevity medicine allowed a valid but easily refutable argument that it might lead to an increase in health inequity. There are abundant reasons to counter-resonate. Firstly,

novel technologies and interventions are developing rapidly in a competitive ecosystem. We have seen a decline in, e.g., genetic testing pricing or CT/MRI imaging. With this trend, the longevity medicine might actually be one of the most affordable domains in terms of diagnostics and follow-up, as well as regarding geroprotective supplements or even specific invasive interventions, such as plasmapheresis. Secondly, most of the hallmarks of aging can be co-influenced by lifestyle modifications (exercise, nutrition, supplements, caloric restriction, intermittent fasting, cognitive activities, etc.), led by inexpensive app-based solutions, such as DACs and CGM [2, 22]. Such management toward risk prevention and short-term improvement of performance does not require strong financial inputs and is accessible to the majority of digitalized populations. An accompanying longevity physician navigating the measurements and interventions is an optimal way to fully extrapolate and exploit these tools, since they are able to customize the interventions and interpret the measurements for an individual, minding the biovariability, comorbidities, chronological age, and – importantly – personal goals and preferences (which determine the compliance). As mentioned previously, these experts are still underrepresented and imply a speedy

advent of appropriate education. Thirdly, healthcare institutions across virtually all countries globally face overcrowding and limited resources due to high healthcare costs, which in return aggravate the health inequality and inequity [50]. Most cost burden derives from age-related and chronic diseases [51, 52]. Mitigating those is now a priority in the medical and scientific field, as to circumvent a clash of health systems challenged with the spiraling chronic multimorbidities. Another example is the simple illustration of cost and time efficacy of AI-based drug discovery and repurposing, which allows to save millions of USD and several years on identification and testing of new compounds. At the moment, the sobering facts are as follows: 90% of all drug trials fail, very few that do succeed take an average of 10 years to reach the market, and cost ranges from \$2.5 to \$12 billion. In addition, in computero clinical trial simulations are promising to bring more equity: algorithms can be trained and retrained to include features that are largely ignored, such as the aspect of biological sex and gender, elderly, multimorbid patients, ethnic minorities, and importantly the age [53, 54].

Surely, as for any emerging field, also precision medicine will face barriers related to socioeconomic inequities and demand solutions toward financial viability, especially in systems based on solidarity. Targeting and empowering coordinated discussions of multidisciplinary stakeholder by continuous updates on the current state of science by KOLs in a transparent manner is a crucial approach toward a successful democratization of longevity medicine for all [55].

Since healthy longevity medicine is precision medicine driven by aging biomarkers and is not seeking a bare extension of life, but an extension of a healthy, productive lifespan, the field can notably contribute to improve the economy of healthcare and as a whole.

AI Applications in Medicine: The Fundament of Precision Medicine

Tangible AI applications in medicine are rapidly increasing in number, complexity, and accuracy, even though overall, they are still in very

preliminary stages [56]. The vast variety of areas and problems that securely designed AI systems in medicine can tackle in practice bear major opportunities toward improvement of the healthcare as entity. At the entrance of the new decade (2020), most AI appliances are targeted at assisting physicians in their diagnoses and treatment decisions [57]. The ultimate goal for the reactive medicine is to achieve an individualized prediction which treatment will work for which patient and how well in order to avoid chronification, physical and emotional burden, and costs. There are several overarching AI appliances that trespass current healthcare. Firstly, the basic prerequisite for AI applications – the data – is now mostly collected in a planned electronic manner. Secondly, AI allowed telemedicine to be established and flourish speedily. Public health impacts are indisputable alone through the provision of medical care in rural regions with a low density of specialists (thus, faster diagnosis, better prognosis, less morbidity and mortality, less costs, less burden on the patients and caregivers, etc.).

AI-based diagnostic systems are able to detect features quickly, quantitatively, objectively, and reproducibly and thus classify conspicuous skin lesions with high accuracy. Machine learning supports diagnostics in numerous specialties, e.g., oncology (lung cancer detection [58, 59]), neurology (CT-assisted stroke detection), ophthalmology (early detection of maculo- and retinopathies [60]), cardiology (risk assessment of myocardial events, sudden cardiac death based on ECG and cardiac MRI), dermatology (detection and classification of dermal lesions based on a body scan or even images [61]), etc. The latter exemplifies the abundant potential of AI applications in patient care, e.g., remote care, patient engagement, and auto-monitoring. Even though the dermatologic smartphone apps for self-examination by patients are still under a strict scrutiny and depend on compliance, pre-identification of suspicious skin lesions by the AI algorithm and a following analysis by specialists improve the early detection of, e.g., melanoma [62]. “AppDoc-Online Dermatologist” and the “Derma-App” are prominent examples in Germany. More studies are needed to confirm the outperformance of AI over dermatologists, based on verified benchmarks [63–65]. A

close follow-up of conspicuous lesions in high-risk patients is further enabled with digital dermoscopy, a 3D whole-body photography [66, 67].

AI can further facilitate and optimize medical processes, such as division and delivery of blood products – a problem that is virtually universal and leads to massive losses in resources (both time and costs). “AutoPiLoT” AI system and app, for example, approaches this issue by evaluating past data and thus is able to make predictions of amount and timing of blood units needed in a specific hospital, as well as to recognize patterns that can further be readjusted individually. The app has now also implemented blood donation, capturing the donor data and thus simplifying the matches [68].

AI can also simplify the diagnostics of complex diseases, through complex diagnostic tools, such as MRI in multiple sclerosis (MS). AI algorithms, encapsulated in a user-friendly app, will allow nonexperts (such as GP) to accompany the patient, interpret complex images over time, and gain experience [69].

Tailoring an optimal treatment for a patient involving all aspects of the clinical picture is considered but optimally also other information about the patient’s biological features (e.g., genetic data), health data, and examination results. Data from a large number of patients must be analyzed and intelligently linked to predict the course of the disease and the optimal therapy for a specific group of patients sharing similar features. AI is incumbent in order to conduct these steps and thus enable what is defined as precision medicine.

Conclusion and Future Perspectives

Longevity medicine is a groundbreaking dynamic field, emerging as one of the most essential medical disciplines combining the most advanced and complex diagnostic and interventional approaches. As an AI-driven precision medicine, longevity medicine harnesses innovative and state-of-the art technologies and science to exploit the potential of the human genome, deep quantitative phenotyping, -omics (e.g., epigenomics, metabolomics, proteomics, etc.), microbiome, radiogenomic precision

imaging, etc. With the help of continued data collection, the development of new features, improvements in ways of interpretation, and further optimizations in the implementation of an individual patient protocol, longevity medicine will be self-perpetuating. The longitudinal approach enables a trifold dynamic, interrelated longevity practice: data mining, patient compliance, and physicians’ lead. This shifts reactive medicine with limited human data analysis capacity toward longevity doctors that can collect and apply gigabytes of patients’ data toward identification, mitigation, and elimination of actionable diseases, preferably years and decades ahead.

References

1. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol.* 2013;14(10):3156.
2. Putin E, Mamoshina P, Aliper A, Korzinkin M, Moskalev A, Kolosov A, et al. Deep biomarkers of human aging: Application of deep neural networks to biomarker development. *Aging* [Internet]. 2016 May 18 [cited 2021 Jan 10];8(5):1021–33. <https://www.aging-us.com/article/100968/text>
3. Melzer D, Pilling LC, Ferrucci L. The genetics of human ageing. *Nat Rev Genet.* 2020;21(2):88–101.
4. Kieft-de Jong JC, Mathers JC, Franco OH. Nutrition and healthy ageing: the key ingredients. *Proc Nutr Soc.* 2014;73(2):249–59.
5. Schehl B, Leukel J. Associations between individual factors, environmental factors, and outdoor independence in older adults. *Eur J Ageing.* 2020;17(3):291–8.
6. Liu Z, Kuo P-L, Horvath S, Crimmins E, Ferrucci L, Levine M. A new aging measure captures morbidity and mortality risk across diverse subpopulations from NHANES IV: a cohort study. *PLoS Med* [Internet]. 2018 Dec 31 [cited 2019 Sep 21];15(12). <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6312200/>
7. Hoshino A, Horvath S, Sridhar A, Chitsazan A, Reh TA. Synchrony and asynchrony between an epigenetic clock and developmental timing. *Sci Rep.* 2019;9(1):1–12.
8. Lu AT, Quach A, Wilson JG, Reiner AP, Aviv A, Raj K, et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (Albany NY).* 2019;11(2):303–27.
9. Bonawitz K, Eichner H, Grieskamp W, Huba D, Ingberman A, Ivanov V, et al. Towards federated learning at scale: system design. *arXiv:190201046 [cs, stat]* [Internet]. 2019 Mar 22 [cited 2021 Jan 10]; <http://arxiv.org/abs/1902.01046>
10. Müezzinler A, Zaineddin AK, Brenner H. A systematic review of leukocyte telomere length and age in adults. *Ageing Res Rev.* 2013;12(2):509–19.

11. Peters MJ, Joehanes R, Pilling LC, Schurmann C, Conneely KN, Powell J, et al. The transcriptional landscape of age in human peripheral blood. *Nat Commun.* 2015;6(1):8570.
12. Mamoshina P, Volosnikova M, Ozerov IV, Putin E, Skibina E, Cortese F, et al. Machine learning on human muscle transcriptomic data for biomarker discovery and tissue-specific drug target identification. *Front Genet [Internet].* 2018 [cited 2021 Jan 10];9. <https://www.frontiersin.org/articles/10.3389/fgene.2018.00242/full>
13. Galkin F, Mamoshina P, Aliper A, de Magalhães JP, Gladyshev VN, Zhavoronkov A. Biohorology and biomarkers of aging: current state-of-the-art, challenges and opportunities. *Ageing Res Rev.* 2020;60:101050.
14. Barzilai N, Crandall JP, Kritchevsky SB, Espeland MA. Metformin as a tool to target aging. *Cell Metab.* 2016;23(6):1060–5.
15. Levin JM, Oprea TI, Davidovich S, Clozel T, Overington JP, Vanhaelen Q, et al. Artificial intelligence, drug repurposing and peer review. *Nat Biotechnol.* 2020;38(10):1127–31.
16. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature.* 2018;562 (7726):203–9.
17. El Emam K, Rodgers S, Malin B. Anonymising and sharing individual patient data. *BMJ [Internet].* 2015 Mar 20 [cited 2021 Jan 10];350. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4707567/>
18. Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. *NPJ Digit Med [Internet].* 2020 Sep 14 [cited 2021 Jan 10];3. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7490367/>
19. Thapa C, Chamikara MAP, Camtepe S. SplitFed: when federated learning meets split learning. *arXiv:200412088 [cs] [Internet].* 2020 Sep 2 [cited 2021 Jan 10]; <http://arxiv.org/abs/2004.12088>
20. Zhang Y, Ling C. A strategy to apply machine learning to small datasets in materials science. *npj Comput Mater.* 2018;4(1):1–8.
21. Radenkovic D, Keogh SB, Maruthappu M. Data science in modern evidence-based medicine. *J R Soc Med.* 2019;112(12):493–4.
22. Zhavoronkov A, Mamoshina P, Vanhaelen Q, Scheibye-Knudsen M, Moskalev A, Aliper A. Artificial intelligence for aging and longevity research: recent advances and perspectives. *Ageing Res Rev.* 2019;49:49–66.
23. Aliper A, Belikov AV, Garazha A, Jellen L, Artemov A, Suntssova M, et al. In search for geroprotectors: in silico screening and in vitro validation of signalome-level mimetics of young healthy state. *Aging (Albany NY).* 2016;8(9):2127–41.
24. NCDs and ageing [Internet]. [cited 2021 Jan 10]. <https://www.who.int/westernpacific/about/governance/regional-director/ncds-and-ageing>
25. Mitchell-Fearon K, Waldron N, Laws H, James K, Holder-Nevins D, Willie-Tyndale D, et al. Non-communicable diseases in an older, aging population: a developing country perspective (Jamaica). *J Health Care Poor Underserved.* 2015;26 (2):475–87.
26. López-Otin C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. *Cell.* 2013;153 (6):1194–217.
27. Deleidi M, Jäggel M, Rubino G. Immune aging, dysmetabolism, and inflammation in neurological diseases. *Front Neurosci.* 2015;9:172.
28. Montecino-Rodriguez E, Berent-Maoz B, Dorshkind K. Causes, consequences, and reversal of immune system aging. *J Clin Invest.* 2013;123(3):958–65.
29. Bruunsgaard H, Pedersen M, Pedersen BK. Aging and proinflammatory cytokines. *Curr Opin Hematol.* 2001;8(3):131–6.
30. Sarkar D, Fisher PB. Molecular mechanisms of aging-associated inflammation. *Cancer Lett.* 2006;236(1): 13–23.
31. Michaud M, Balardy L, Moulis G, Gaudin C, Peyrot C, Vellas B, et al. Proinflammatory cytokines, aging, and age-related diseases. *J Am Med Dir Assoc.* 2013;14 (12):877–82.
32. Franceschi C, Capri M, Monti D, Giunta S, Olivieri F, Sevini F, et al. Inflammaging and anti-inflammaging: a systemic perspective on aging and longevity emerged from studies in humans. *Mech Ageing Dev.* 2007;128 (1):92–105.
33. Kapetanaki MG, Mora AL, Rojas M. Influence of age on wound healing and fibrosis. *J Pathol.* 2013;229(2): 310–22.
34. Cieslik KA, Taffet GE, Carlson S, Hermosillo J, Trial J, Entman ML. Immune-inflammatory dysregulation modulates the incidence of progressive fibrosis and diastolic stiffness in the aging heart. *J Mol Cell Cardiol.* 2011;50(1):248–56.
35. Shindyapina AV, Mkrtchyan GV, Gneteeva T, Buiuci S, Tancowny B, Kulka M, et al. Mineralization of the connective tissue: a complex molecular process leading to age-related loss of function. *Rejuvenation Res.* 2013;17(2):116–33.
36. van Deursen JM. The role of senescent cells in ageing. *Nature.* 2014;509(7501):439–46.
37. Moskalev AA, Aliper AM, Smit-McBride Z, Buzdin A, Zhavoronkov A. Genetics and epigenetics of aging and longevity. *Cell Cycle.* 2014;13(7):1063–77.
38. Lombard DB, Chua KF, Mostoslavsky R, Franco S, Gostissa M, Alt FW. DNA repair, genome stability, and aging. *Cell.* 2005;120(4):497–512.
39. Lardenoije R, Iatrou A, Kenis G, Komplotis K, Steinbusch HWM, Mastroeni D, et al. The epigenetics of aging and neurodegeneration. *Prog Neurobiol.* 2015;131:21–64.
40. Helling BA, Yang IV. Epigenetics in lung fibrosis: from pathobiology to treatment perspective. *Curr Opin Pulm Med.* 2015;21(5):454–62.
41. De Rosa M, Pace U, Rega D, Costabile V, Duraturo F, Izzo P, et al. Genetics, diagnosis and management of colorectal cancer (Review). *Oncol Rep.* 2015;34(3): 1087–96.
42. Aguilar-Olivos NE, Oria-Hernández J, Ponciano-Rodríguez G, Chávez-Tapia NC, Uribe M, Méndez-

- Sánchez N. The role of epigenetics in the progression of non-alcoholic fatty liver disease. *Mini Rev Med Chem.* 2015;15(14):1187–94.
43. Kennedy BK, Berger SL, Brunet A, Campisi J, Cuervo AM, Epel ES, et al. Geroscience: linking aging to chronic disease. *Cell.* 2014;159(4):709–13.
44. Zhavoronkov A, Buzdin AA, Garazha AV, Borisov NM, Moskalev AA. Signaling pathway cloud regulation for in silico screening and ranking of the potential geroprotective drugs. *Front Genet [Internet].* 2014 Mar 3 [cited 2021 Jan 10];5. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3940060/>
45. Blagosklonny MV. Validation of anti-aging drugs by treating age-related diseases. *Aging (Albany NY).* 2009;1(3):281–8.
46. Gaweijn E, Hiss JA, Schneider G. Deep learning in drug discovery. *Mol Inform.* 2016;35(1):3–14.
47. Mincholé A, Rodriguez B. Artificial intelligence for the electrocardiogram. *Nat Med.* 2019;25(1):22–3.
48. Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. *Mol Syst Biol.* 2016;12(7):878.
49. Mamoshina P, Vieira A, Putin E, Zhavoronkov A. Applications of deep learning in biomedicine. *Mol Pharm.* 2016;13(5):1445–54.
50. Kruk ME, Gage AD, Arsenault C, Jordan K, Leslie HH, Roder-DeWan S, et al. High-quality health systems in the Sustainable Development Goals era: time for a revolution. *Lancet Glob Health.* 2018;6(11):e1196–252.
51. Lancet T. Tackling the burden of chronic diseases in the USA. *Lancet.* 2009;373(9659):185.
52. Soriano JB, Kendrick PJ, Paulson KR, Gupta V, Abrams EM, Adedoyin RA, et al. Prevalence and attributable health burden of chronic respiratory diseases, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet Respir Med.* 2020;8(6):585–96.
53. Mamoshina P, Bueno-Orovio A, Rodriguez B. Dual transcriptomic and molecular machine learning predicts all major clinical forms of drug cardiotoxicity. *Front Pharmacol [Internet].* 2020 [cited 2021 Jan 10];11. <https://www.frontiersin.org/articles/10.3389/fphar.2020.00639/full>
54. Bakula D, Aliper AM, Mamoshina P, Petr MA, Teklu A, Baur JA, et al. Aging and drug discovery. *Aging (Albany NY).* 2018;10(11):3079–88.
55. Fohner AE, Volk KG, Woodahl EL. Democratizing precision medicine through community engagement. *Clin Pharmacol Ther.* 2019;106(3):488–90.
56. Amisha MP, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. *J Fam Med Prim Care.* 2019;8(7):2328–31.
57. Sloane EB, J. Silva R. Artificial intelligence in medical devices and clinical decision support systems. In: Clinical engineering handbook. Academic; 2020. p. 556–68.
58. Sathyakumar K, Munoz M, Singh J, Hussain N, Babu BA. Automated lung cancer detection using artificial intelligence (AI) deep convolutional neural networks: a narrative literature review. *Cureus [Internet].* [cited 2021 Feb 15];12(8). <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7518939/>
59. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med.* 2019;25(6):954–61.
60. Ruamviboonsuk P, Krause J, Chotcomwongse P, Sayres R, Raman R, Widner K, et al. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. *npj Digit Med.* 2019;2(1):1–9.
61. Dick V, Sinz C, Mittlböck M, Kittler H, Tschandl P. Accuracy of computer-aided diagnosis of melanoma: a Meta-analysis. *JAMA Dermatol.* 2019;155(11):1291.
62. Cui X, Wei R, Gong L, Qi R, Zhao Z, Chen H, et al. Assessing the effectiveness of artificial intelligence methods for melanoma: a retrospective review. *J Am Acad Dermatol.* 2019;81(5):1176–80.
63. Brinker TJ, Hekler A, Hauschild A, Berking C, Schilling B, Enk AH, et al. Comparing artificial intelligence algorithms to 157 German dermatologists: the melanoma classification benchmark. *Eur J Cancer.* 2019;111:30–7.
64. Aractingi S, Pellacani G. Computational neural network in melanocytic lesions diagnosis: artificial intelligence to improve diagnosis in dermatology? *Eur J Dermatol.* 2019;29(S1):4–7.
65. Brinker TJ, Hekler A, Enk AH, Berking C, Haferkamp S, Hauschild A, et al. Deep neural networks are superior to dermatologists in melanoma image classification. *Eur J Cancer.* 2019;119:11–7.
66. Rubegni P, Burroni M, Perotti R, Fimiani M, Andreassi L, Cevenini G, et al. Digital dermoscopy analysis and artificial neural network for the differentiation of clinically atypical pigmented skin lesions: a retrospective study. *J Investig Dermatol.* 2002;119(2):471–4.
67. Gewirtzman AJ, Braun RP. Computerized digital dermoscopy. *J Cosmet Dermatol.* 2003;2(1):14–20.
68. Ärzteblatt DÄG Redaktion Deutsches. Künstliche Intelligenz soll für bessere Verteilung von Blutkonserven... [Internet]. Deutsches Ärzteblatt. 2021 [cited 2021 Feb 15]. <https://www.aerzteblatt.de/nachrichten/119899/Kuenstliche-Intelligenz-soll-fuer-bessere-Verteilung-von-Blutkonserven-sorgen>
69. Afzal HMR, Luo S, Ramadan S, Lechner-Scott J. The emerging role of artificial intelligence in multiple sclerosis imaging. *Mult Scler.* 2020. 1352458520966298.



Joseph Davids and Hutan Ashrafiyan

Contents

Introduction	1170
Review of Literature	1172
Results and Discussion	1172
Machine Learning for Nanodrug Discovery and Nanoformulations	1175
Drug Delivery Systems and Formulation	1175
Machine Learning in Specific Areas of Nanomedicine	1175
Mathematical Machine Learning Modeling for Cancer Nanomedicine and Theranostics	1175
Machine Learning in Precision Nanotheranostics for Cancer	1179
Machine Learning for Nanotoxicology	1180
Machine Learning and Quantum Enabled Technologies	1181
Machine Learning and Regenerative Nanobiology	1181

J. Davids (✉)
Imperial College London NHS Trust, London, UK

Hamlyn Centre for Robotics and Artificial Intelligence,
Department of Surgery and Cancer, Imperial College
London, London, UK

National Hospital for Neurology and Neurosurgery Queen
Square, London, UK
e-mail: j davids@ic.ac.uk

H. Ashrafiyan
Department of Surgery and Cancer, Imperial College
London NHS Trust, London, UK

Institute of Global Health Innovation, Imperial College
London, London, UK

Hamlyn Centre for Robotics and Artificial Intelligence,
Department of Surgery and Cancer, Imperial College
London, London, UK

Next Generation Machine Learning for Nanorobotic Surgery	1181
Ethics and Regulation	1182
References	1183

Abstract

Nanotechnology and its sister field of quantum technologies are interdisciplinary sciences that have been touted as one of the holy grails of technological advancements still yet to reach critical mass and unveil their transformative potential. Similarly, artificial intelligence and machine learning constitute another technological advancement that has captivated scientific hearts and minds, with both leading to next generation industrial revolutions.

The unification of the latter and former technologies has thus elevated opportunities for exciting emerging discoveries and promises to offer further combinatorically exponential translational discoveries for medicine and humankind. This chapter explores the use of AI in the subfield of nanomedicine. We explore the applications of machine learning algorithms to aspects of drug discovery, toxicology and regenerative medicine, as well as medical and surgical robotics.

Keywords

AI in nanomedicine · Nanotechnology · Nanoethics · Nano-neuroscience · Quantum surgery · Nanorobotic · Nanoformulation · Drug delivery · Nanocarrier · Gold nanoparticles

Introduction

Nanotechnology and its sister fields of quantum technologies have delivered on their promise as exciting scientific disciplines continuing to revolutionize our understanding of the atomic-scale properties of matter. Here we adopt the European Commission's definition of "nano" at 1–100 nm for >50% of the number size distribution of particles sizes, excluding materials with surface area

by volume in excess of $60 \text{ m}^2/\text{cm}^3$ [1]. With respect to environmental health and safety or pollution-related concerns, this number size distribution threshold of 50% may be substituted by a threshold between 1 and 50% [1]. A formal definition of nanomedicine is the use of nanomaterials for the diagnosis, monitoring, control, prevention, and treatment of disease [2]. Nanomedicine thus leverages nanomaterials for medical disease theranostics (a combination of diagnosis and the simultaneous potential for therapeutics) and has allowed remarkable medical nanomaterial discoveries to be made. The fabrication of these nanomaterials is either in a top-down or a bottom-up approach [1]. Either method influences the material properties and as such the constraints that must be considered before construction and modeling of the material is achieved. However, this does allow for a multitude of possibilities designing different types of materials from their bulk substrates.

On the nanoscale, the significant characteristics that allow the properties to differ from the bulk material are determined by the materials' intrinsic and physiognomically determined properties. These properties include: size, shape, surface curvature, and chemical composition, as well as properties such as crystallinity, porosity, particle heterogeneity, roughness, zeta potential or surface charge—determining hydrophobicity and hydrophilicity, and conferred ability for functionalization, etc. Toxicological properties include particle aggregation, dissolution features, the ability of the material to be in a state of dispersion, which aids its solubility, the ionic strength, pH, temperature, and the internal surrounding environment that the particle is interacting with, which may have the presence of large organic molecules (for example, proteins) as well as the external environment after the particle's life cycle and hence biodegradability.

The general principles of nanotechnology for medicine thus include [3]:

- Size and environmental interactions
- Self-assembly and pseudo-intelligence
- Feedback control
- Biomimicry
- Molecular machines
- Cross-collaborative multidisciplinary approaches such as from biology, engineering, chemistry, physics, and medicine

It is unsurprising that several of these described characteristics of nanomaterials can

be studied and predicted using applied machine learning and artificial intelligence methodologies. For a complete treatment of artificial intelligence, we refer you to the respective chapters, but have summarized the process in Fig. 1. However, this chapter will look at areas where machine learning techniques have been leveraged to help resolve complexities associated with nanomaterial design and drug discovery in nanomedicine. The literature and argument for machine learning methodology in nanomedical

Machine Learning Cycle

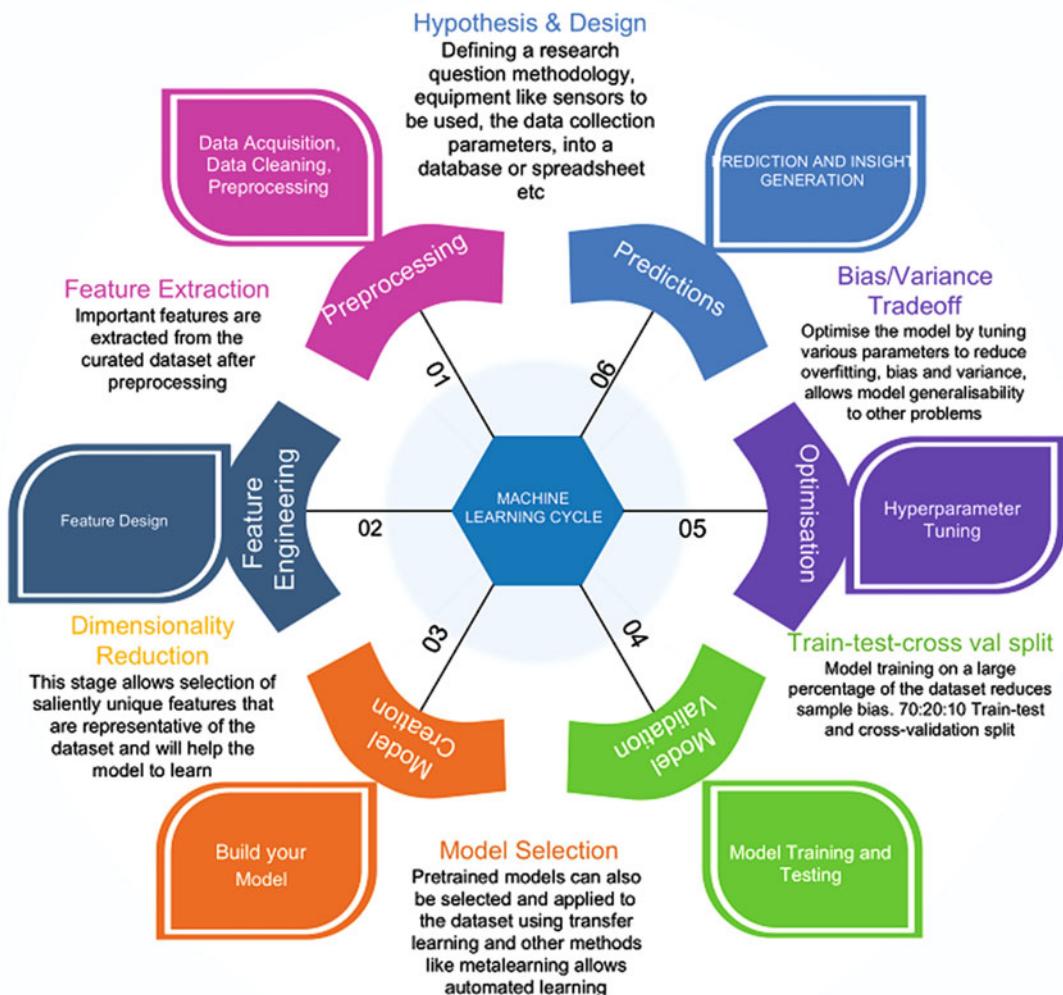


Fig. 1 A representational example of the machine learning cycle. Methodologies vary, but the underlying central tenets

remain ubiquitous. The cyclical nature is seen whereby there is transition from one step to the next in an iterative fashion

discovery is still embryonic, compounded by the argument that the study and fabrication process is usually probabilistic. This makes model explainability challenging with some machine learning methods, such as deep learning, very difficult due to its “blackbox” processes. Nonetheless there are various areas where successful inroads have been made with machine learning, including in the fields of nanotoxicology, nano-oncology, DNA and other genetic nanotechnologies and regenerative nanomedicine, etc. Nanoparticles can augment cancer drug delivery to increase efficacy with a wide range of formulated customizations available to counteract cancer drug resistance, immune cloning for reduced drug clearance and drug distribution, etc. Identifying these unique customizations will benefit from methods that can allow combinatorically unique feature and entity extraction to aid drug design [4].

The chapter begins with a review of the literature in this area and tackles various aspects of the application of artificial intelligence to facilitate nanomedicine discovery. We also discuss nanobots in medicine, their inherent intelligences, and the challenges faced in terms of control. We then conclude our discussions looking at some of the ethical and regulatory aspects of nanomedicine discovery and applications.

Review of Literature

We searched the following databases:

- Books@Ovid <January 19, 2021>
- Journals@Ovid Full Text <February 10, 2021>
- AMED (Allied and Complementary Medicine) <1985 to January 2021>
- Embase Classic+Embase <1947 to 2021 February 11>
- Ovid Emcare <1995 to 2021 Week 05>
- Global Health Archive <1910 to 1972

- Health and Psychosocial Instruments <1985 to October 2020>
- HMIC Health Management Information Consortium <1979 to November 2020>
- Ovid MEDLINE(R) ALL <1946 to February 10, 2021>
- Maternity and Infant Care Database (MIDIRS) <1971 to December 2020>
- PUBMED

using the following search terms including truncations:

- “Machine Learning” and Nanomed\$.mp. [mp = tx, bt, ti, ab, ct, sh, hw, tn, ot, dm, mf, dv, kw, fx, dq, id, cc, ac, de, md, sd, ip, vo, pg, pu, ib, is, yr, et, jn, dp, ja, bd, ar, bs, cf, pj, pa, dt, rf, so, pb, nm, kf, ox, px, rx, an, ui, ds, on, sy], generated 154 results
- “Artificial Intelligence” and Nanomed\$.mp. [mp = tx, bt, ti, ab, ct, sh, hw, tn, ot, dm, mf, dv, kw, fx, dq, id, cc, ac, de, md, sd, ip, vo, pg, pu, ib, is, yr, et, jn, dp, ja, bd, ar, bs, cf, pj, pa, dt, rf, so, pb, nm, kf, ox, px, rx, an, ui, ds, on, sy], generated 139 results

Results and Discussion

The literature from the above databases identified 293 results in total, suggesting that this is still an area that will continue to grow in the coming years, with the required further research investment, and that it has the potential to deliver innovative approaches. Table 1 summarizes the results of AI models used in nanomedicine research through additional citation tracking and analysis of review articles.

Recent decades have seen an explosion of interest in nanobiointerface technologies, which have their origins deeply rooted in nanomaterial engineering principles and are also coupled with various biological entities such as extracellular matrix scaffolds. Nanoparticles can interact with DNA,

Table 1 Machine learning models identified in literature for nanomedicine

Author	Type of nanomaterial studied	Tissue of interest	Machine learning model proposed	Reported model accuracy
Ban et al. 2020 [5]	Multiwalled and single-walled carbon nanotubes, metallic (e.g., Ag, Au, and Fe ₃ O ₄), nonmetallic (e.g., SiO ₂ and Si), and liposomal nanoparticles (e.g., cholesterol-phosphatidylcholine and thiolated amino-poly[ethylene glycol]; 3 kDa) N-acetyl-L-cysteine and thiolated L-asparagine), cationic (e.g., 11-amino-1-undeca- nethiol and hexadecyltrimethylammonium bromide), neutral (e.g., carboxymethyl-poly[ethylene glycol]-thiol [5 kDa] and bicyclononyne), common (e.g., carboxyl [COOH] and citrate [CIT]), and rare (e.g., Pluronic F-127 and 16-mercaptopentadecanoic acid) surface ligands	Plasma from healthy samples, serum, blood, or cell culture medium	Random forest, neural networks, and linear regression	R ² = 0.80
Basso et al. 2021 [6]	Cationic lipid, glycerol-based lipids, GLY1, and GLY2 conjugates Drugs included: calcium, atorvastatin, and curcumin	Brain tumor cells	Unsupervised: Row-wise principal component analysis, hierarchical clustering analysis Supervised: PLS regression with noniterative partial least squares, neural networks with TanH linear and Gaussian activation functions	R ² = 0.95–0.99
Sayes and Ivanov [7]	Metal oxide nanoparticles, titanium oxide	Rat alveolar macrophages and immortalized cell lines studying LDH release	Linear discriminant analysis for classification	R ² = 0.77
Puzyn et al. [8]	Metal oxide nanoparticles (ZnO, CuO, V ₂ O ₃ , Y ₂ O ₃ , Bi ₂ O ₃ , In ₂ O ₃ , Sb ₂ O ₃ , Al ₂ O ₃ , Fe ₂ O ₃ , SiO ₂ , ZrO ₂ , SnO ₂ , TiO ₂ , CoO, NiO, Cr ₂ O ₃ , and La ₂ O ₃)	<i>E. Coli</i> bacteria looking at EC ₅₀	Multiple regression	R ² = 0.85
Liu et al. [9]	Metal oxide nanoparticles (Al ₂ O ₃ , CeO ₂ , Co ₃ O ₄ , TiO ₂ , ZnO, CuO, SiO ₂ , Fe ₃ O ₄ , and WO ₃)	BEAS-2B looking at membrane permeability	Logistic regression models	Accuracy = 100%

(continued)

Table 1 (continued)

Author	Type of nanomaterial studied	Tissue of interest	Machine learning model proposed	Reported model accuracy
Horev-Azaria et al. [10]	Metal oxide nanoparticle (CoFe_2O_4)	A549, NCI H441, HepG2, MDCK, Caco-2 TC7, TK6, and primary mouse dendritic cells Cell viability assessment	J48	Accuracy = 92.5%
Winkler et al. [11]	Metal oxide nanoparticle (CLIO)	Endothelial muscle cells, smooth muscle cells, hepatocytes, and monocytes assessing smooth muscle apoptosis	Bayesian neural networks and multilinear regression models	$R^2 = 0.90$
Fourches et al. [12]	Metal nanoparticles (CLIO, pseudocaged, monocrystalline iron oxide, CdSe core quantum dot, and iron based)	Biological activity profiles of nanoparticles in monocytes, hepatocytes, endothelial cells, and smooth muscle cells	Support vector machines	Accuracy = 88%
Toropova et al. [13]	Metal oxide nanoparticles	pLC_{50} studied in <i>E. coli</i>	Monte Carlo simulation	$R^2 = 0.9835$
Jones et al. [14]	PAMAM dendrimers	Caco-2 cell viability studies	J48	Accuracy = 83.5%
Liu et al. [15]	Metal nanoparticles, dendrimer, metal oxide, and polymeric materials	24-h post-fertilization mortality in zebrafish	IBK	Accuracy = 83.7%
Shalaby et al. [16]	Di- and triblock copolymers of polyethylene glycol and polylactide	Looking at particle size and entrapment	Artificial neural network	$R^2 = 0.96484$
Davids and Carlisle [17]	Gold nanoparticles	Looking at the formation of gold nanoparticles of various sizes	Deep neural networks	Accuracy = 93.85%

Modified from Jones et al. [9, 14, 17, 18]

proteins, cell membrane lipid bilayers, and various intracellular organelles, such as endosomal compartments, to establish a series of connected interfaces between the material and cellular substrate [19].

This interaction enhances the biophysico-chemical, thermodynamic, and kinetic properties. Example interactions include protein corona formation and intracellular uptake. The design of these nanobiointerfaces by interfacial chemists, physicists, engineers, and medics to help study and resolve various diseases necessitate a combinatorial approach. A few

approaches in the past decades have leveraged combinatorial methods to help decipher the complexity, variability, and interspecies differences of these biointerfaces in colloidal form. The boundaries that shape the interfaces themselves encompass three aspects: (1) the nanoparticle surfaces that have features derived from their unique size and shape-determined physicochemical properties, which have been modeled using combinatorial machine learning; (2) the contact zone between the particle and its surrounding environment or media before it is

interfaced with the biological tissue surface; and (3) the zone of interactivity, where the nanomaterial comes into contact with the biological surface, can also be modeled combinatorically [19]. However, the challenge of computational complexity limits the resolution. The process is also computationally resource intensive for dynamic modeling to be effective. Here too applications of machine learning approaches, such as generative adversarial networks and in particular quantum machine learning, could augment the process for building nanobiointerfaces through intelligent material design by resolving molecule-molecule interactions. The behavior of these interfaces, surface chemistry, and how they interact with each other and other molecules within the body are areas that will continue to benefit from other machine learning approaches for which deep learning will prove useful.

It is imperative that force and bonding interaction modeling such as hydrogen bonding, Van der Waals, and ionic interactions be achieved using quantum machine learning because of the probabilistic quantum characteristics of the electron cloud behavior necessitating more advanced methods of data capture and processing. Quantum computing and advances in this space are allowing modeling of hydrogen bonding to be fully appreciated and studied at better atomic resolutions. The next section describes some of these machine learning models in specific areas of nanomedicine. We begin with drug discovery and nanoformulation.

Machine Learning for Nanodrug Discovery and Nanoformulations

Drug Delivery Systems and Formulation

Various approaches are adopted to ensure that the designed nanodrugs are delivered in a timely manner. One approach includes modifications to the pharmacokinetics and pharmacodynamics to improve patient compliance while sustaining optimal treatment [20]. Contrary to

immediate release preparations, modified release preparations can also be designed that allow effective maintenance of the drug concentrations within acceptable limits to counteract the side effect profile and also ensure appropriately consistent drug dosing. These approaches are achieved through a variety of technologies including liposomal formulations, microsphere platforms, nanoparticle suspension conjugates, or nanoemulsions [20]. For an in-depth review of the approaches used please see the detailed work by Cottura et al. and with a summary of their findings in Table 2 below [20]. This area has benefited from mathematical modeling using QSAR and nano-QSAR for predicting nanoparticle to cellular interactions such as toxicity, cellular uptake, drug loading and release profile, and nanoparticle receptor binding characteristics [8, 12, 18]. Current approaches that also utilize artificial intelligence through reinforcement learning for computational decision-making in drug discovery and nanoformulation have been discussed by Vamathevan et al. [21]

Machine Learning in Specific Areas of Nanomedicine

Mathematical Machine Learning Modeling for Cancer Nanomedicine and Theranostics

Cancer remains a very challenging area of research with rapid advancements and discovery on a daily basis. Understanding cancer data is paramount for modeling and drug discovery [24]. In silico models offer an opportunity for deriving insights into cancer drug discovery. However, the vast dataset and heterogeneity makes the description of information on cancer drug development and progression at times uncorrelated [25]. Many researchers therefore continue to seek new big data-driven and machine learning in silico modeling techniques that correlate with *in vivo* data for cancer drug discovery. [25]

Drugs such as contrast agents and RNA silencing therapies all require a nanocarrier-based design to carry the drug to the required location

Table 2 Summary of pharmaceutical nanoformulations that have been adopted to augment drug delivery

Delivery system	Drug	Active ingredient	Composition	Company	Indication	Approval date
Liposome	Doxil ^a	Doxorubicin hydrochloride	HSPC and PEG	Janssen	Ovarian cancer, sarcoma, myeloma	1995
Ambisome	Amphotericin B	HSPC and DSPG	Astellas	Fungal infection	1997	
Depocyt	Cytarabine	DOPC and DPPG	Pacira	Lymphomatous	1999	
Epxarel	Bupivacaine	DOPC and DOPE	Pacira	Local anesthetic	2011	
Marqibo kit	Vincristine sulfate	Eggs sphingomyelin	Talon	Acute lymphoblastic leukemia	2012	
Onivyde	Imiquimod hydrochloride	DSPC and MPEG-2000-DSPE	Ipsen	Adenocarcinoma of the pancreas	2015	
Microsphere	Lupron Depot	Leuprolide acetate	PLGA	Abbvie	Advanced prostatic cancer	1995
Sandostatin LAR	Octreotide acetate	PLGA	Novartis	Acromegaly	1998	
Trelistar	Triptorelin pamoate	PLGA	Allergen	Advanced prostate cancer	2000	
Definity	Perflutren	DPPA, DPPC, and MPEG-5000-DPPE	Lanthus	Ultrasound contrast agent	2001	
Risperdal Consta	Risperidone	PLG	Janssen	Schizophrenia, bipolar I disorder	2003	
Vivitrol	Naltrexone	PLG	Alkermes	Alcohol dependence	2006	
Bydureon	Exenatide synthetic	PLGA	AstraZeneca AB	Type 2 diabetes	2012	
Signifor lar	Pasireotide pamoate	PLGA	Novartis	Acromegaly	2014	
Lumason	Sulfur hexafluoride Lipid-type microspheres	DSPC	Bracco	Ultrasound contrast agent	2014	
Bydureon BCise	Exenatide	PLGA	AstraZeneca AB	Type 2 diabetes	2017	
Triptodur kit	Triptorelin pamoate	PLGA	Arbor	Central precocious puberty	2017	
Bicillin L-A	Benzathine penicillin	Dispersion	King Pharma	Rheumatic fever	1952	
Suspension and nanoparticle	Depo-Provera	MPA	Dispersion (microcrystalline suspension)	Pharmacia & Upjohn	Contraception	1959
Atridox	Doxycycline hyclate	PLA	Tolmar	Chronic adult periodontitis	1998	
Eligard	Leuproreotide acetate	PLGA	Tolmar	Advanced prostate cancer	2002	

	<i>Abraxane</i> ^b	<i>Paclitaxel</i>	<i>Protein nanoparticle</i>	<i>Abraxis</i>	<i>Metastatic breast cancer; non-small cell lung cancer</i>	2005
Somatuline Depot	Lantreotide acetate	Nanotide		Ipsen	Acromegaly	2007
Zyprexa Relپrevv	Olanzapine pamoate	Microcrystal	Eli Lilly	Schizophrenia		2009
Invega Sustenna	Paliperidone palmitate	Nanocrystal	Janssen	Schizophrenia		2009
Feraheme	Ferumoxytol	Carbohydrate-coated iron-oxide nanoparticle	Amag	Iron deficiency anaemia		2009
Abilify Maintena	Aripiprazole	Nanocrystal	Osuka	Schizophrenia		2013
Ryanodex	Dantrolene sodium	Nanocrystal	Eagle	Malignant hyperthermia		2014
Invega Trinza (3-month)	Paliperidone palmitate	Nanocrystal	Janssen	Schizophrenia		2015
Aristada Sustol	Aripiprazole lauroxil Granisetron	Nanocrystal Ortho ester	Alkermes Heron	Schizophrenia Nausea and vomiting		2015 2016
Sublocade	Buprenorphine	PLGA	Indivior	Moderate to severe opioid use disorder		2017
Perseris Cabenuva	Risperidone Rilpivirine	In situ forming gel Dispersion	Indivior Janssen	Schizophrenia HIV		2018 Clinical trials
Emulsion Intralipid	Soybean oil	Fat emulsion	Fresenius	Parenteral nutrition		1975
Haldol Cleviprex	Haloperidol decanoate Clevidipine	Oil depot Lipid emulsion	King Pharma Chiesi	Schizophrenia Reduction of blood pressure		2000 2008
Smoflupid	Fish oil	Lipid emulsion	Frerenius	Parenteral nutrition		2016
Cinvanti	Aprepitant	Lipid emulsion	Heron	Acute and delayed nausea and vomiting		2017

HIV, human immune deficiency virus; DOPC, dioleyoylphosphatidylcholine; DOPE, dioleoylphosphatidylethanolamine; DPPG, dipalmitoylphosphatidylglycerol; DSPC, distearoylphosphatidylcholine; DSPE, distearoylphosphatidylethanolamine; DSPG, distearoylphosphatidylglycerol; HSPC, hydrogenated soyphosphatidylcholine; MPA, mycophenolic acid; MPEG, methoxypolyethylene glycol; NA, not applicable; PEG, polyethylene glyco; PLA, polyactide; PLG(A), poly(lactic-co-glycolic [acid])

^aFirst US Food and Drug Administration approved nanodrug

^bFirst US Food and Drug Administration approved nanotechnology-based target drug delivery

Adapted from Cottura et al. [20, 22, 23]

using stealth-cloaking systems to bypass the body's immune system and avoid premature clearance [20, 26]. The voyage and fate of the nanodrug from entry into the body is a: (1) vascular, (2) trans-vascular, (3) interstitial, or (4) a pinocytic-clathrin-mediated transport allowing inspissation into cells or a diffusion-based trans-cellular mechanism [20, 26–28].

Immediately after injection, they encounter a high proportion (60–80 g/L) of plasma proteins, such as albumin, and phagocytes, opsonins, and lipopolysaccharides that they have to avoid by means of adsorption, which transforms the biochemical and molecular identity and fingerprint of the particle by creating biomolecular corona around it [20, 28]. This biomolecular corona has hydrodynamic capabilities giving it a newer bioidentity. Augmented bioidentities affect charge-charge interactions, enterohepatic and hemodynamic interactions determined by changing capillary network pore sizes, which affect the particle-to-pore size ratio, hydrophobicity, and pH changes that the drug will encounter on its journey to the destination of interest. These can be modeled using machine learning and big data approaches to allow effective predictions for the newly fabricated nanodrug.

When at its destination, an effective design must allow the carrier system to exploit the characteristic biophysicochemical milieu of the tumor microenvironment to jettison its payload to the tissue of interest. Through a phenomenon known as margination, the nanoparticle that has now gained the ability to enter the cell accumulates at the microvasculature cell wall boundary [29].

After it has performed its desired effect, toxicity limiting measures must be built into the nanocarrier and nanodrug to allow their removal or clearance to minimize damage to the convalescing normal tissue. The interaction of nanodrugs with both noncancer and cancer cells has thus gained considerable research focus with significant rigor.

The drugs can accumulate through passive accumulation, ionic binding, or Van der Waals interactions depending on the ability to functionalize the nanoparticle with another bioconjugate, which causes pharmacokinetic and pharmacodynamic modifications [20, 26, 28]. This often leads to half-life changes, triggered release by

enzymatic or electromagnetic interaction, temperature, and pH-dependent changes when the environment interacts with the nanodrug. These drugs can now be fabricated as organic, polymeric, or liposomal drugs or inorganic, gold, silica, diamondoid, and carbon-based and even DNA/RNA/aptamer-based constructs all with unique characteristics that lend themselves to added interaction complexities, which the nanomedic must appreciate and learn to exploit or overcome. Such combinatorial vastness lends itself well to artificial intelligence and big data solutions to help overcome complex interaction analysis. What remains intriguing from various cancer nanomedicine meta-analyses is the fact that on average only 0.7% of the drug reaches the destination tumor [20, 25, 28, 30].

Consequently, areas such as enhanced permeability retention characteristics or leakiness of the tumor microvasculature, and interactions of these microenvironments with nanochemotherapy drugs require augmented methods to advance discovery of the features that lead to suboptimal nanocarrier drug delivery. Machine learning is one such area that promises a supporting role for effective mathematical modeling within this space.

In silico models of the entry of the particle into the body include kinetic and thermodynamic models and molecular dynamics simulations, all of which can be theoretically achieved with machine learning. Particle transport and extravasation into the targeted tissue environment have been modeled using mesoscopic length scale, continuum, and discrete modeling. Machine learning models like random forests have also been used to predict the corona formation [5]. In Ban et al's study, corona compositions classified by physicochemical and functional properties such as length, theoretical isoelectric point (pI), molecular weight, grand average of hydropathicity (GRAVY) score, and function were utilized. We describe another example model below for corona formation [5]. Dogra et al. provide a spatiotemporal classification of mathematical models for nanomedicine cancer therapeutics: (1) convection, diffusion, reaction kinetics, PKPD models, (2) continuum discrete hybrid models, and (3) agent-based models that enable molecular dynamic simulations to be generated [28].

Modeling the protein nanoparticle interactions with cancer cell receptors was proposed by Dell'Orco and colleagues who derived the equations below [31].

$$\frac{d[\text{Nanoparticle} \times \text{Protein}]}{dt} = n_{\text{protein}} k_{\text{protein}}^{\text{on}} [\text{Nanoparticle}] [\text{Protein}] - k_{\text{protein}}^{\text{off}} [\text{Nanoparticle} \times \text{Protein}] \quad (1)$$

where

$[\text{Nanoparticle}]$ = Concentration of nanoparticles

$[\text{Proteins}]$ = Concentration of proteins

$k_{\text{protein}}^{\text{off}}$ = Off-rate constants when there is non-bound protein to receptor

$k_{\text{protein}}^{\text{on}}$ = On-rate constant when there is protein bound to receptor

$$n_{\text{protein}} = \frac{4\pi(r_{\text{nano}} + r_{\text{protein}})^2}{\pi r_{\text{protein}}^2} \quad (2)$$

where n_{protein} is the number of available binding sites on the nanoparticle surface, r is the radius of either the nanoparticle and the protein (assuming that the protein is a sphere) to give a total cross-sectional area for the protein surface.

This was extended by Darabi Sahneh and colleagues to include the concentration of available binding sites [32].

$$\frac{d[\text{Nanoparticle} \times \text{Protein}]}{dt} = n_{\text{protein}} k_{\text{protein}}^{\text{on}} [\text{Protein}] Y - k_{\text{protein}}^{\text{off}} [\text{Nanoparticle} \times \text{Protein}] \quad (3)$$

$$\text{where } Y = \left([\text{Nanoparticles}]_0 - \sum_{j=1}^m \frac{[\text{Protein}_j \times \text{Nanoparticle}]}{n_{\text{protein}_j}} \right) \quad (4)$$

The dynamic evolution of the corona formation is thus captured. Physically, force modeling including drag coefficients could also be included to augment these models, but may be deemed

negligible under certain conditions for model simplification. Theoretically these models can form templates for dynamic time series-related pharmacokinetic machine learning model design using real-time sensors to capture the data, but in practice other methods may be preferred. The models are summarized in Table 3, but for an extended review the reader is directed to Dogra and colleagues [28]. The above metrics can be encoded as unique features and fed to a quantum deep learning algorithm to enable predictions for how the particle journeys and interacts with protein receptors as well as its clearance from the body. We are likely to continue to see this more and more in coming years.

Other mathematical models like quantitative structure–activity relationships (QSAR) and their nanoscale equivalent (nano-QSAR) are heavily used for nanotoxicity predictions and are discussed in the section below [40].

Machine Learning in Precision Nanotheranostics for Cancer

In brain tumor research for a safer alternative to a cationic compound for nanodrug delivery, glycerol-based lipids with alkylated chains and glycerol backbones were evaluated using supervised and unsupervised machine learning techniques to identify structural similarities in GLY1 and GLY2 nanolipid carriers, which differ in their polarity, and also to assess the hemocompatibility profile. They showed that GLY1 nanocarriers demonstrated better zeta potential profile with decreasing particle size and also demonstrated reduced cytotoxicity [6, 41].

In renal adenocarcinoma, current AI platforms have focused on optimization for chemotherapeutic drug combinations. Most have been designed to streamline feedback controls that optimize combinatorial drug variations to inhibit the viability of renal adenocarcinoma cell line 786-O [42]. Others have aimed to also optimize nano-diamond to chemotherapy drug combinations (e.g., nano-diamond-doxorubicin conjugate, nano-diamond-mitoxantrone, and nano-diamond-bleomycin) for the best cancer cytotoxicity

Table 3 A nonexhaustive summary of mathematical models that can be augmented using machine learning for cancer theranostics and drug discovery adapted from Dogra et al. [28]

Author	Mathematical model presented	Area of interest	Results described and application
Dell'Orco et al. [31] Darabi Sahneh et al. [32]	Kinetic models see Eqs. 1, 2, 3, and 4	Biomolecular corona formation	Law of mass action, nanoparticle to protein interaction. Series of first-order differential equations modeling the protein concentrations, concentration of protein to nanoparticles, association and dissociation constants. Corona formation. Evolution of metastable to stable state
Lopez and Lobaskin [33]	Coarse-grained molecular dynamic simulations	Biomolecular corona formation	Characteristic protein adsorption energies were modeled and simulated. Modeling results revealed the size-dependent nanoparticle to drug protein receptor binding rather than charge
Gentile et al. [34]	Continuum modeling	Microvascular transport, margination, and binding	Hematocrit and vessel permeability-related reduction in nanoparticle diffusion coefficient
Zhdanov, and Cho [35]	Augmented kinetic models	Adsorption and desorption of proteins	Diffusion rate constant drops at the colloid solid interface
Huajian et al. [36]	Discrete modeling	Cellular internalization of the nanoparticle	There is a threshold minimal particle size and ligand density needed for endocytosis-mediated transport
Dogra et al. [37]	Pharmacokinetic modeling	Whole-body biodistribution and clearance	The smaller the size of the particles the longer they stay in the circulation and are able to evade the immune system. Excretion is achieved by positively charging the nanoparticle. Cloaking reduces their vulnerability to sequestration
Chauhan et al. [38]	Hybrid modeling	Drug delivery into the tumor	Size and other physical and chemical characteristics affect interactions with blood vessels which in turn has a consistent influence on the enhanced permeability retention properties
Pascal et al. [39]	Pharmacodynamic modeling	Therapeutic efficacy and toxicity models	Dose and integration time-dependent effect on how toxic the nanoparticle can be when it comes into contact with the tissue of interest

with a minimal toxicity profile for the surrounding tissues [42]. Nano-diamonds have superior infratoxicity profiles compared to other nano-conjugated and bulk chemotherapeutics and optimization enables rapid toxicology profiling and also can make them likely candidates for computer vision applications [43]. Other noteworthy nano-related platforms designed for uro-oncology include the Quadratic Phenotype Optimization Platform identifying the optimal combination from a pool of 114 FDA approved drugs, including dactinomycin and decitabine [42].

In metastatic castration resistant prostate cancer theranostics, CURATE.AI was used for a guided combination therapy consisting of ZEN-3694 (a bromodomain and extraterminal inhibitor) and enzalutamide (an androgen receptor

inhibitor) [42]. The CURATE.AI was leveraged for patient treatment response monitoring by enabling cautious dose modifications to the ZEN-3694 and enzalutamide chemotherapy drugs. The CURATE.AI-guided treatment was shown to reduce prostate-specific antigen levels and stopped disease progression.

Machine Learning for Nanotoxicology

In nanotoxicology, *in silico* models are predominantly developed for predicting the toxicity of nanoparticles and their interactions with various tissue constructs. The quest to design safer and more cost-effective nanoparticles for nanomedicine has led to a reliance on machine learning and data

mining for drug discovery [40]. Structure-based mathematical models remain the most reported and heavily studied for nanomaterial-induced toxicity, and include quantitative structure–activity relationships (QSAR) and their nanoscale equivalent (nano-QSAR), which attempt to model and predict physicochemically defined, biofunctional molecular properties. These properties include point of zero zeta potentials predicting agglomeration behavior [40].

Other structural-based machine learning models also include Bayesian methods and Markov Chain Monte Carlo simulation. Other models discussed include quantitative nanostructure activity relationship (QNAR) models [20]. What is clear is that different species can have different drug absorption pharmacokinetic features that need to be modeled differently.

Table 1 provides a summary of artificial intelligence and machine learning methods that have been used for various nanotoxicology studies. Most reported studies on toxicity demonstrate predilection toward the use of metal oxide nanoparticles for theranostics. Cytotoxicity from polyamidoamine dendrimers, which are useful delivery systems for nanodrug delivery, were identified using a data mining approach with an accuracy ranging from 60.2–83.5%. Jones and colleagues built a decision tree that indicated that if the isoelectric point was greater than 12.63, dendrimer toxicity was predicted [14]. The predictive toxicology of cobalt nanoparticles in various cell lines representing the lungs, liver, kidney, intestines, and immune system was compared with their ionic forms using a J48 decision tree machine learning model [44]. Their model predicted concentration as the highest rank determinant for toxicity, with a hierarchy of cell sensitivity toward cobalt ions. They also concluded that the toxicity profile of aggregated cobalt nanoparticles was related to their dissolved ionic forms.

Machine Learning and Quantum Enabled Technologies

The purported meteoric rise of quantum computers from works of fiction to big research laboratories and now to their near decentralized

distribution into smaller institutions has allowed quantum machine learning algorithms to permeate into recent consumer testing platforms [45]. IBM, Google, Microsoft, and many others have all invested heavily for quantum supremacy [46]. Quantum computers and the machine learning algorithms developed from them are allowing simulation at atomic-scale resolution for biomolecular interactions, biochemical reaction modeling, and personalised genomic patient data discovery with the promise of reduced drug toxicity and precision medicine and surgery [45]. Current NeuraLink technologies explore these quantum concepts in computational chip design for brain machine interfaces in the quest for human-machine symbiosis [47].

Machine Learning and Regenerative Nanobiology

Regenerative medicine promises to enable treatment of neurodegenerative disease, facilitate new organ regeneration for transplantation, and achieve better tissue scaffold constructions. It works through reprogramming the body's stem cells or inducing pluripotency and direct lineage differentiation to replace an injured or diseased cell [48]. Current augmentations with CRISPR-CAS-9 genomic engineering and gene editing technologies hope to evolve the process even further [49].

Next Generation Machine Learning for Nanorobotic Surgery

A progressive review by Ashrafiyan et al. discusses the evolution and various generations of robotics and the cost implications for surgery and how this will impact anesthesia in the future [50]. Their insight into fifth generation autonomy with human consciousness and the fact that this still remains conceptual is worth noting. As such, we will not place too much emphasis on surgical anesthesia here. However, we will discuss potential future applications of nanorobotics in ophthalmology and neurosurgery, including areas such as neurosurgery and nano-neurosurgical theranostics

[51]. We will also briefly consider conceptually the feasibility of quantum neurosurgery and programmable matter.

In ophthalmology, experimental nanorobotic agents with propellers have been designed and proven to be able to navigate through vitreous humor to facilitate drug delivery or ophthalmic surgery and treat age-related macular degeneration [52].

Neurosurgery also involves leveraging microscopic techniques to allow micrometer gains in accuracy to be achieved. The surgical corridors are usually very much narrower and restrictive compared to other surgeries. As a consequence, the operative risks and margin of error are significantly elevated compared with other surgical specialties. Precision is vital and as a result the increased adoption of minimally invasive methods leveraging microscopy have become the gold standard for most subspecialties within the field [53]. The proximity of eloquent structures to lesions has to be accounted for and tractography is sometimes considered for connectivity-related data discovery [54]. Nano-neurosurgery and quantum neurosurgery purport to offer an even more elevated level of precision surgery, reducing or even eliminating the margin of error to negligible levels. The future of neurosurgery will be augmented by tele-nano-neurorobotic platforms that will enable theranostic efficacies far beyond what is available now, together with facilitating long-distance surgery. This would be possible due to advances in quantum computational speedups, which will help overcome the Brownian challenge of controlling nanoscale devices. As such, enhanced master-slave capabilities and autonomous nanorobotic variations will become available provided the institutional and global cost-benefit analyses are favorable. Patients and clinicians will both have a panoply of choice for targeted therapies.

The design of the operative microscope will also continue to evolve into modular intelligent nanoscopes with femto-level resolution, capable of resolving cellular-to-cellular and subatomic interactions in diseased tissue. The combination of nanorobots and advanced AI algorithms *in vivo* would facilitate programmable matter during in

situ surgery. This will be useful for dynamic real-time cancer nanoablative and precision nano-lesioning surgery, augmented intracellular organelle probing for nano-neurogenetic surgery, reduced surgical complication, and personalized patient-specific nanosurgical presimulation. At these considerable subatomic-scale resolutions, quantum algorithms will be required to stabilize the systems and enable dynamic image processing. There is considerable potential for most of these advancements to no longer be just theoretical or hype in the coming decades. Although realistically it will be dependent upon the surgical need and the preference of the surgeon to adopt newer techniques, which will also need to go through the necessary ethical and regulatory rigor before patients are put at risk.

Ethics and Regulation

The reader is referred to various discussions of AI and ethics in the relevant chapters. However, we will briefly mention some of the issues here. The FDA only recently formally suggested a definition for nanomaterials or nanomedicines issuing only a draft guidance, but no regulatory definition has been established [1].

"The FDA has not established regulatory definitions of "nanotechnology," "nanomaterial," "nanoscale," or other related terms...when considering whether an FDA-regulated product involves the application of nanotechnology, FDA will ask: (1) whether a material or end product is engineered to have at least one external dimension, or an internal or surface structure, in the nanoscale range (approximately 1 nm to 100 nm) and (2) whether a material or end product is engineered to exhibit properties or phenomena, including physical or chemical properties or biological effects, that are attributable to its dimension(s), even if these dimensions fall outside the nanoscale range, up to one micrometre (1,000 nm)." [55]

This consequently creates controversy as to the design of nanomedicines, as there is a lack of regulatory clarity, global consensus view, or established ethicolegal frameworks.

Appreciably, even a change in shape or size might affect its physicochemical interactions

within the body and could lead to altered toxicological profile. Machine learning algorithms should therefore be primed with strict definitions for nanomaterials in their predictions for drug design to ensure that they do conform to regulations and safety parameters.

Machine learning algorithms are only recently starting to also gain regulatory scrutiny as medical devices [56]. They published a discussion paper in 2019 about a proposed regulatory framework for modifications to artificial intelligence–/machine learning–based software as a medical device [56]. MHRA have also had discussions for the regulatory impacts for artificial intelligence for safety monitoring and diagnostic services, but have yet to discuss how the unity of these two technologies would be scrutinized [57]. Regulatory sand-boxing to test these are no doubt in progress [58]. A recent document outlines the three main regulatory areas of concern and include the following [59]:

- Level of autonomy introduced by AI technologies
- Ability of continuous learning systems to change their output over time in response to new data
- Ability to explain and understand how an output has been reached

As regards the nanoethical perspective of nanomedicines, the fact that we still do not fully understand the toxicity profiles of most nanomedicines and nanomaterials make it a problem for new material development and design applications to be translated to patients. It can also take several decades from conception to clinical trials for eventual treatments to be released, which is too costly a risk for the medical and pharmaceutical stakeholders and makes them apprehensive. So, the addition of another black box process such as artificial intelligence and specifically deep learning, where model explainability is necessary, ethically compounds this problem even further. However, as hippocratic physicians, the casuist viewpoint still holds that we have a moral obligation to develop newer and more effective treatment modalities

using artificial intelligence and nanotechnologies to minimize harm in medicine.

References

1. Soares S, Sousa J, Pais A, Vitorino C. Nanomedicine: principles, properties, and regulatory issues. *Front Chem.* 2018;6:360. <https://doi.org/10.3389/fchem.2018.00360>. PMID: 30177965; PMCID: PMC6109690
2. Tinkle SMS, Muhlebach S, et al. Nanomedicines: addressing the scientific and regulatory gap. *Ann N Y Acad Sci.* 2014;1313:35–56.
3. Zarbin M, Montemagno C, Leary J, Ritch R. Nanotechnology in ophthalmology a b. *Can J Ophthalmol.* 2010;45:457–76.
4. Balaz I, Petric T, Kovacevic M, Tsompanas M-A, Stillman N. Harnessing adaptive novelty for automated generation of cancer treatments. *Bio Systems.* 2021;199:104290.
5. Ban Z, Yuan P, Yu F, Peng T, Zhou Q, Hu X. Machine learning predicts the functional composition of the protein corona and the cellular recognition of nanoparticles. *Proc Natl Acad Sci U S A.* 2020;117(19):10492–9.
6. Basso J, Mendes M, Silva J, Cova T, Luque-Michel E, Jorge AF, et al. Sorting hidden patterns in nanoparticle performance for glioblastoma using machine learning algorithms. *Int J Pharm.* 2021;592:120095.
7. Sayes C, Ivanov I. Comparative study of predictive computational models for nanoparticle-induced cytotoxicity. *Risk Anal.* 2010;30(11):1723–34. Epub 2010/06/22
8. Puzyn T, Rasulev B, Gajewicz A, Hu X, Dasari TP, Michalkova A, et al. Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. *Nat Nanotechnol.* 2011;6(3):175–8. Epub 2011/02/15
9. Liu R, Rallo R, George S, Ji Z, Nair S, Nel AE, et al. Classification NanoSAR development for cytotoxicity of metal oxide nanoparticles. *Small.* 2011;7(8):1118–26. Epub 2011/04/02
10. Horev-Azaria L, Baldi G, Beno D, Bonacchi D, Golla-Schindler U, Kirkpatrick JC, et al. Predictive toxicology of cobalt ferrite nanoparticles: comparative in-vitro study of different cellular models using methods of knowledge discovery from data. *Part Fibre Toxicol.* 2013;10:32. Epub 2013/07/31
11. Winkler D, Burden FR, Yan B, Weissleder R, Tassa C, Shaw S, et al. Modelling and predicting the biological effects of nanomaterials. *SAR QSAR Environ Res.* 2014;25(2):161–72. Epub 2014/03/15
12. Fourches D, Pu D, Tassa C, Weissleder R, Shaw SY, Mumper RJ, et al. Quantitative nanostructure-activity relationship modeling. *ACS Nano.* 2010;4(10):5703–12. Epub 2010/09/23
13. Toropova A, Toropov AA, Rallo R, Leszczynska D, Leszczynski J. Optimal descriptor as a translator of

- eclectic data into prediction of cytotoxicity for metal oxide nanoparticles under different conditions. *Ecotoxicol Environ Saf.* 2015;112:39–45. Epub 2014/12/03.
14. Jones DE, Ghandehari H, Facelli JC. Predicting cytotoxicity of PAMAM dendrimers using molecular descriptors. *Beilstein J Nanotechnol.* 2015;6:1886–96.
 15. Liu X, Tang K, Harper S, Harper B, Steevens JA, Xu R. Predictive modeling of nanomaterial exposure effects in biological systems. *Int J Nanomedicine.* 2013;8(Suppl 1):31–43. Epub 2013/10/08.
 16. Shalaby K, Soliman ME, Casettari L, Bonacucina G, Cespi M, Palmieri GF, et al. Determination of factors controlling the particle size and entrapment efficiency of noscapine in PEG/PLA nanoparticles using artificial neural networks. *Int J Nanomedicine.* 2014;9:4953–64. Epub 2014/11/05.
 17. Davids J, Carlisle, R. Artificial nano-intelligence: using deep learning models to study the formation of gold nanoparticles: potential conceptual applications of transfer learning in the field of nano-neuroscience drug discovery In: Frontiers, editor. Neural bases of action – from cellular microcircuits to large-scale networks and modelling; erice: Frontiers Neuroscience; 2018.
 18. Jones DE, Ghandehari H, Facelli JC. A review of the applications of data mining and machine learning for the prediction of biomedical properties of nanoparticles. *Comput Methods Prog Biomed.* 2016;132:93–103.
 19. Nel AE, Madler L, Velegol D, Xia T, Hoek EM, Somasundaran P, et al. Understanding biophysicochemical interactions at the nano-bio interface. *Nat Mater.* 2009;8(7):543–57.
 20. Cottura N, Howarth A, Rajoli R, Sicardi M. The current landscape of novel formulations and the role of mathematical modeling in their development. *J Clin Pharmacol.* 2020;60(S1):S77–97.
 21. Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov.* 2019;18(6):463–77.
 22. FDA. U. DOXIL. <https://www.accessdata.fda.gov/scripts/cder/daf/index.cfm?event=overview.process&ApplNo=050718>. Accessed 27 Feb 2021.
 23. FDA. U. Ambisome. https://www.accessdata.fda.gov/drugsatfda_docs/nda/97/050740_ambisome_toc.cfm. Accessed 27 February 2021.
 24. Thomas DG, Klaessig F, Harper SL, Fritts M, Hoover MD, Gaheen S, et al. Informatics and standards for nanomedicine technology. *Wiley Interdiscip Rev Nanomed Nanobiotechnol.* 2011;3(5):511–32.
 25. Cova TFGG, Bento DJ, Nunes SCC. Computational approaches in theranostics: mining and predicting cancer data. *Pharmaceutics.* 2019;11(3):119.
 26. Oh JY, Kim HS, Palanikumar L, et al. Cloaking nanoparticles with protein corona shield for targeted drug delivery. *Nat Commun.* 2018;9:4548. <https://doi.org/10.1038/s41467-018-06979-4>.
 27. Kaksonen M, Roux A. Mechanisms of clathrin-mediated endocytosis. *Nat Rev Mol Cell Biol.* 2018;19:313–26. <https://doi.org/10.1038/nrm.2017.132>.
 28. Dogra P, Butner JD, Chuang YL, Caserta S, Goel S, Brinker CJ, et al. Mathematical modeling in cancer nanomedicine: a review. *Biomed Microdevices.* 2019;21(2):40.
 29. Müller K, Fedosov D, Gompper G. Margination of micro- and nano-particles in blood flow and its effect on drug delivery. *Sci Rep.* 2014;4:4871. <https://doi.org/10.1038/srep04871>.
 30. Thakur V, Kutty RV. Recent advances in nanotheranostics for triple negative breast cancer treatment. *J Exp Clin Cancer Res.* 2019;38(1):430.
 31. Dell'Orco DLM, Oslakovic C, Cedervall T, Linse S. Modeling the time evolution of the nanoparticle-protein corona in a body fluid. *PLoS One.* 2010;5(6):e10949. <https://doi.org/10.1371/journal.pone.0010949>.
 32. Darabi Sahneh FSC, Riviere J. Dynamics of nanoparticle-protein corona complex formation: analytical results from population Balance equations. *PLoS One.* 8(5):e64690. <https://doi.org/10.1371/journal.pone.0064690>.
 33. Lopez Hal V. Coarse-grained model of adsorption of blood plasma proteins onto nanoparticles. *J Chem Phys.* 2015;143:243138. <https://doi.org/10.1063/1.4936908>.
 34. Gentile F, Ferrari M, Decuzzi P. The transport of nanoparticles in blood vessels: the effect of vessel permeability and blood rheology. *Ann Biomed Eng.* 2008;36:254–61. <https://doi.org/10.1007/s10439-007-9423-6>.
 35. Zhdanov V, Cho NJ. Kinetics of the formation of a protein corona around nanoparticles. *Math Biosci.* 2016;282:82–90.
 36. Huajian G, Shi W, Freund LB. Mechanics of receptor-mediated endocytosis. *Proc Natl Acad Sci.* 2005;102:9469–74.
 37. Dogra P, Adolphi NL, Wang Z, Lin YS, Butler KS, Durfee PN, Croissant JG, Noureddine A, Coker EN, Bearer EL, Cristini V, Jeffrey Brinker C. Establishing the effects of mesoporous silica nanoparticle properties on *in vivo* disposition using imaging-based pharmacokinetics. *Nat Commun.* 2018;9:4551.
 38. Chauhan V, Stylianopoulos T, Martin JD, Popović Z, Chen O, Kamoun WS, Bawendi MG, Fukumura D, Jain RK. Normalization of tumour blood vessels improves the delivery of nanomedicines in a size-dependent manner. *Nat Nanotechnol.* 2012;7:383–8.
 39. Pascal J, Ashley CE, Wang Z, Brocato TA, Butner JD, Carnes EC, Koay EJ, Brinker CJ, Cristini V. Mechanistic modeling identifies drug-uptake history as predictor of tumor drug resistance and nano-carrier-mediated response. *ACS Nano.* 2013;7(12):11174–82.
 40. Singh AV, Ansari MHD, Rosenkranz D, Maharjan RS, Kriegel FL, Gandhi K, et al. Artificial intelligence and machine learning in computational Nanotoxicology: unlocking and empowering nanomedicine. *Adv Healthc Mater.* 2020;9(17):e1901862.
 41. Agrahari V, Burnouf P-A, Burnouf T, Agrahari V. Nanoformulation properties, characterization, and behavior in complex biological matrices: challenges and opportunities for brain-targeted drug delivery

- applications and enhanced translational potential. *Adv Drug Deliv Rev.* 2019;148:146–80.
42. Wilson B, Km G. Artificial intelligence and related technologies enabled nanomedicine for advanced cancer treatment. *Nanomedicine (London).* 2020;15(5):433–5.
43. Yu S, Kang M, Chang H, Chen K, Yu Y. Bright fluorescent nanodiamonds: no photobleaching and low cytotoxicity. *J Am Chem Soc.* 2005;127:17604–5.
44. Horev-Azaria L, Kirkpatrick C, Korenstein R, Marche P, Maimon O, Ponti J, et al. Predictive toxicology of cobalt nanoparticles and ions: comparative in vitro study of different cellular models using methods of knowledge discovery from data. *Toxicol Sci.* 2011;122(2):489–501.
45. Hassanzadeh P. Towards the quantum-enabled technologies for development of drugs or delivery systems. *J Control Release.* 2020;324:260–79.
46. Arute F, Arya K, Babbush R, Bacon D, Bardin JC, Barends R, et al. Quantum supremacy using a programmable superconducting processor. *Nature.* 2019;574 (7779):505–10.
47. Musk E, Neuralink, Shukla H. An integrated brain-machine interface platform with thousands of channels. *J Med Internet Res* 2019;21(10):e16194.
48. Izadifar M. An artificial intelligence approach to develop tunable nanoparticulate delivery systems for regenerative medicine applications. *BioImpacts.* 2018;8(Supplement 1):12. <https://doi.org/10.1517/bi.2018.S1>. Available from: <https://www.fda.gov/media/88828/download>.
49. Ran F, Hsu P, Wright J, et al. Genome engineering using the CRISPR-Cas9 system. *Nat Protoc.* 2013;8:2281–308. <https://doi.org/10.1038/nprot.2013.143>.
50. Ashrafian H, Clancy O, Grover V, Darzi A. The evolution of robotic surgery: surgical and anaesthetic aspects. *BJA.* 2017;119(suppl_1):i72–84. <https://doi.org/10.1093/bja/aex383>.
51. Kateb B, Heiss J. Final cover_The TextBook of Nano-neuroscience and Nanoneurosurgery_K122372013.
52. Wu Z, Troll J, Jeong H, Wei Q, Stang M, Ziemssen F, Wang Z, Dong M, Schnichels S, Qiu T, Fischer P. A swarm of slippery micropropellers penetrates the vitreous body of the eye. *Sci Adv.* 2018;4(11):eaat4388. <https://doi.org/10.1126/sciadv.aat4388>. Available from: <https://www.fda.gov/media/88828/download>.
53. Liu C, Spicer M, Apuzzo M. The genesis of neurosurgery and the evolution of the neurosurgical operative environment: part II-concepts for future development, 2003 and beyond. *Neurosurgery.* 2003;52(1):20–35.
54. Akram H, Dayal V, Mahlknecht P, Georgiev D, Hyam J, Foltyne T, et al. Connectivity derived thalamic segmentation in deep brain stimulation for tremor. *NeuroImage: Clin.* 2018;18:130–42.
55. FDA. Use of nanomaterials in food for animals. US Department of Health and Human Services Food and Drug Administration Center for Veterinary Medicine 2015;August 2015.
56. FDA. Artificial intelligence and machine learning in software as a medical device: federal food and drug administration; 2021. Available from: <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device>.
57. MHRA. MHRA role in development and use of artificial intelligence for safety monitoring: A board meeting. 2018. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/749139/Item_07_42-OB-2018_Artificial_Intelligence.pdf.
58. CQC. Using machine learning in diagnostic services- a report with recommendations from CQC's regulatory sandbox. 2020. Available from: https://www.cqc.org.uk/sites/default/files/20200324%20CQC%20sandbox%20report_machine%20learning%20in%20diagnostic%20services.pdf.
59. BSI, MHRA, AAMI. The emergence of artificial intelligence and machine learning algorithms in healthcare: recommendations to support governance and regulation. 2019. Available from: <https://www.bsigroup.com/globalassets/localfiles/en-gb/about-bsi/nsb/innovation/mhra-ai-paper-2019.pdf>.



AIM in Wearable and Implantable Computing

85

Annalisa Baronetto and Oliver Amft

Contents

Introduction	1188
Context Awareness	1189
Definition and Theory	1190
Context Types and Recognition Path	1191
Selected Recognition Problems	1192
Context Pattern Spotting and Interpretation	1192
Interpretation from Context Hierarchy	1193
Flu Detection	1193
Situation Interpretation in Implants	1194
Design and Construction	1194
Topology Design and Optimization	1195
Wearable Personalization	1195
Prefabrication and Process Planning	1196
AI-Assisted Fabrication of Wearables	1196
Validation	1197
AI System safety assessment	1197
Implantable System Testing	1198
Self-Checking AI Systems	1199
Usability	1199
References	1199

Abstract

Wearable and implantable computing devices are becoming an integral component in healthcare thanks to their ability to measure multiple physical, physiological, and environmental variables in everyday life. The analysis of the large volume, velocity, and variety of data requires AI methods and context awareness. Body-worn computing devices can help clinicians by providing important information in

A. Baronetto
Digital Health, FAU Erlangen-Nürnberg, Erlangen,
Germany
e-mail: annalisa.baronetto@fau.de

O. Amft (✉)
Digital Health, Friedrich-Alexander University Erlangen-Nuremberg, Erlangen, Germany
e-mail: amft@computer.org; oliver.amft@fau.de

decision-making, e.g., during diagnosis and therapy management, but similarly could help to implement interventions, e.g., to provide reminders and nudges. This chapter gives an overview on AI-empowered wearable and implantable computing systems and their use in medicine. First, context awareness and its implementation is introduced. AI application examples are detailed. Subsequently, AI methods for personalization and optimal system design are reviewed, including methods for fabrication planning and monitoring. The chapter is concluded by discussing procedures to validate medical wearable and implantable systems, from safety assessments to usability tests.

Keywords

Wearable computing · Medical wearables · Ubiquitous computing · Edge intelligence · Edge computing · Smart textiles · Smart clothing · Embedded systems

Abbreviations

AM	Additive Manufacturing
ANN	Artificial Neural Network
APM	Autonomous pacemakers
CAD	Computer-Aided-Design
CGM	Continuous Glucose Monitoring
CNN	Convolutional Neural Network
ECG	Electrocardiography
GAN	Generative Adversarial Network
HMM	Hidden Markov Model
ICD	Implantable Cardioverter Defibrillator
IMU	Inertial Sensor Unit
IoT	Internet of Things
MCDM	Multi-Criteria Decision Making
ML	Machine Learning
SVM	Support Vector Machine
SVT	Supraventricular tachycardia
VT	Ventricular tachycardia

Introduction

Wearable and implantable devices can collect vital signs, behavior, and exposure data in everyday life outside of controlled clinical settings. The volume, velocity, and variety of collected data cannot be examined manually by clinicians

[1]. Instead, AI methods and Machine Learning (ML) models can be adapted to automatically recognize relevant patterns or statistically summarizes data recorded by worn or implanted sensors as well as interaction with the devices. AI can be integrated on devices to achieve one of the following tasks [2]:

1. Help preventing chronic conditions by predicting the prognosis.
2. Support the diagnosis based on retrieved information. In addition, devices can improve existing screening protocol for early disease detection.
3. Monitor patient parameters and detect anomalies to improve the therapy management.

A key driver of wearable and implantable computing is to manage individual patient needs and to optimize patient-related outcomes by integrating data of the environment (exposome) and behavior (behaviorome) into the care process [1]. In fact, close-to body monitoring data is a basic requirement for personalized prevention and care, as the environmental and behavioral data is not accessible to classic clinical analyses. The rapid adoption of smartphone apps for specific medical needs – even with insufficient clinical evidence – demonstrates, personalizing technology has the potential to revolutionize medical care. Initiatives on personalized or precision medicine specifically consider data-driven technologies for monitoring, as well as diagnostic and treatment support [3]. However, systematic reviews in digital health on the benefit of medical and health apps show that at present many have a small effect [4]. While there are various medical applications, where the functions of smartphones can be leveraged for patient interaction, there are many areas where dedicated or local sensors and actuators are meaningful or essential. An example on the benefit of wearables are the SARS-CoV-2 inspired initiatives to track infection symptoms: The German “Corona Datenspende” has excited more than 500k users to donate smartwatch data including resting heart rate and step count, which may indicate fever, e.g., resting heart rate increase by 8.5 beats per minute per additional 1°K body

temperature, details in [5]. Similar efforts to monitor the pandemic are underway elsewhere. Besides the ability of wearables and implants to produce big data, the tracking initiatives show that the devices provide data several days faster than institutional influenza reports.

Advances in embedded systems have been driven by technology miniaturization, digital sensors, and more recently, by algorithms. In particular, wearable and implantable systems are benefitting from the advances and systematically integrate all key computing components. Data input come from a combination of sensor and user interaction channels, powerful and efficient microcontroller architectures provide comprehensive resources, displays, actuators, data storage, low-power communication interfaces round-off the feature set. The convergence in embedded computing can be observed from the extended function and instruction sets offered by modern microcontrollers: On the peripheral side, a variety of digital busses and analog interfaces have become a standard. At the computing core, microcontrollers offer signal processing functions, e.g. compute the Fast Fourier Transform, or even programmable logic, besides standard instruction sets and integrated memory. Wearable and implantable systems build on the concepts of edge computing and internet-of-things (IoT) [6], i.e. to integrate and aggregate data locally as far as possible, which necessitates further controller features that support AI algorithms, including functions for large matrix operation and deep network processing. Empowered by the advanced computing around the body, users and developers find themselves confronted with a wide application and solution space. However, it is similarly important to note that electronics miniaturization has slowed down since the 2010s already, which brings resource efficiency, personalized design, and material technology into the focus of further research and development efforts [7]. An overview on the design challenges of wearable computers [8] details placement-related wearable design considerations related among others to wearability, robustness, social acceptance, safety, and cost.

An important long-term trend for wearable and implantable computing systems is that the

boundaries of on-body vs. in-body technology get blurred. Whereas with the first wave of biosensors in the 1970s [9], recent monitoring devices may be on-body but half-implanted, such as hearing aids and earables, e.g. [10], or contact lenses, e.g. [11]. For example, recent advances in continuous glucose monitoring (CGM) have yielded on-body, “insertable” monitoring devices that maintain a subcutaneous sensor, i.e. an electrode under the skin, while protecting the compromised tissue layers from infections by a wearable patch [12, 13]. For specific insulin bolus calculations, the CGM measurements alone are insufficient. Additional data from user interaction or sensors, e.g. for physical activity, are needed for optimal bolus estimation. Consequently, CGM insertables may communicate readings or estimates to the users smartphone that serves as data integration and analytics hub for those other sensors too. The focus of this chapter is on the AI-related data and computing, e.g. data abstraction and pattern recognition, performed at the wearable and implantable device, thus mostly disregarding the mobile phones.

The chapter is structured as follows. Section “[Context Awareness](#)” focuses on context awareness in wearable and implantable systems, i.e. methods to recognize the situation of the device and user as well as to derive digital biomarkers. Section “[Selected Recognition Problems](#)” presents selected application examples of context-aware systems and discussed their AI methods as well as open challenges. Subsequently, section “[Design and Construction](#)” illustrates AI methods for designing body-worn devices and section “[Validation](#)” deals with the device and algorithm validation.

Context Awareness

This section provides a concise overview on the principles of context awareness and their implementation in wearable and implantable systems. Starting from the problem framing and definitions, a typical AI processing path is discussed. Subsequently, the context types and their timing are introduced before providing examples of algorithms and applications.

Definition and Theory

Context awareness deals with the computer representation of any given situation of the device or system, the user, and their environment. Often context is considered from a user perspective, i.e. representing the user's state. There have been several definitions of context offered over time that emphasize different aspects of the situation representation [14, 15]. For example, in their project survey, Chen and Kotz [16] refer to environmental states and settings that determine application behavior or events that are interesting to the user. In 2001, Dey et al. [17] described context as any information that helps characterizing the situation of entities. Derived from the definitions of context, an application or system is context aware when it uses context information to filter information and/or adjust the application behavior. Dey et al. offer an overview on conceptual components and architecture of context aware systems, including sensors, widgets, interpreters, aggregators, services, and discoverers.

- **Sensors:** Provide data input, e.g. from measurements of location.
- **Widgets:** Hide the complexity of the sensor source, e.g. whether an radio-frequency indoor positioning system is used or a sensor floor, and abstract context information for applications. Widgets are providing context information in the form needed by application, e.g. room changes, not fine-grained position changes within a room. They serve as reusable building blocks for applications.
- **Interpreters:** Raise abstraction of context information, e.g. by AI-based inference. Similar to the concept of a widget, interpreters are building blocks that applications may use.
- **Aggregators:** Collect multiple pieces of context information. Aggregators are needed when context information is generated in a distributed form, e.g. by different widgets. In addition, the component is used to infer context based on the information available at an aggregator.
- **Services:** Execute actions on behalf of an application. Services are responsible for controlling or changing a state, e.g. using an

actuator. Similar to widgets for the sensor side, services abstract the complexity inherent to changing context state.

- **Discoverers:** Maintain a registry of available widgets, interpreters, aggregators, and services. In particular for wearable and implantable systems, certain widgets and services may be only temporary available due to changes in the environment, location, or resources. For example, during low energy states, some services may become unavailable.

In their 2016 survey, Yürür et al. [18] follow a conceptualization related to Dey et al., but generalize sensors to *context sources* that may reflect the nature of virtual sensors, e.g. activity state of a user-activated panel or screen and information sourced implicitly from user interaction. Aggregators relate to *context managers* that integrate context sources and an adaptation manager to query, process, and filter all context information. Yürür et al. denote that all described processing components are part of a context-aware middleware platform that resides between data sources and the context-aware application.

Context awareness has fundamental challenges related to the interpretation and representation of the world in any computer. Assuming that the world can be represented by a random process of ideal internal states, it is immediately evident that context must be a subset of those states, which are actually observable. In other words, internal states may remain hidden, if there is no mapping function found to relate the internal states to observable ones. The lack of mapping functions is the first fundamental challenge of context awareness and the related context recognition.

Another fundamental challenge of context awareness is related to observations. Assuming that a state-dependent random process of all potential observations exists, measurements denote a subset of all possible observations given the technically available means. Thus the second challenge of context awareness is that available observations and measurements are sparse, e.g. discrete time samples, and noisy, e.g. limited by the performance of the measurement device.

In summary, context awareness in computing systems is an inverse, underdetermined problem, i.e. starting from sparse measurements to represent an internal world state. Thus, a global context recognition is intractable. Most context recognition thus builds on assumptions, which may imply bias of the developer or the underlying data.

Context Types and Recognition Path

Already in 1994, Schilit et al. [14] described the basic relevance of context aware applications that may build on mobile computing, e.g. for offering location-dependent services, but are not limited to those. In addition, the authors emphasize that context information includes further aspects, i.e. who you are with and what resources are nearby. In more formal terms, they refer to environments, including the computing environment (available processing resources), the user environment (location, social situation), and the physical environment (lighting and noise level). Yürür et al. [18] proposed a hierarchical definition of the context representation into low-level (physical, virtual, and logical sensors), high-level (context of device and user, physical and temporal context), and situational relationships, which is referred to as user state.

From the medical care perspective, any context representation is part of either the environment (exposome) or the behavior (behaviorome) [1]. In a user-centered context typing approach for wearable and implantable medical systems, we extend the representation from Lukowicz et al. [19] as follows:

- **User activity:** Physical activities, e.g. walking, eating, but including those activities related to social interaction too, e.g. speaking, listening. Measurements may include body motion, user utterances.
- **User state:** Physiological and cognitive state, e.g. muscle fatigue, mental arousal, arterial fibrillation, which may reflect physical effort too. Clearly, many disease symptoms are reflected in the user state and thus render user state estimation as an essential feature of

context awareness. Measurements may include heart and respiratory rate.

- **Environmental state:** Physical status of the user's surrounding, e.g. rainfall, open window. Measurements may include light and sound.
- **Location:** Navigation and position data related to the user's spatial point, which may be indoors, e.g. in the office, in an elevator, or outdoors, e.g. at the train station. Measurements may include wireless signal strength from hotspots or in-building beacons, as well as global positioning system data.
- **Digital activity:** Activity in digital and social media, e.g. website visits and blog posts. Measurements may include number of posts and likes.
- **Background data:** Subsumed information, typically in a statistical representation, of historic data. Data examples include a patient's electronic health record, frequency statistics from questionnaires, and similar. Background data could be understood as a prior for an AI-based inference system and may be used explicitly by the available data (see above examples) or implicitly by assumption or design (e.g. informative priors or expert knowledge).

While context types provide a spatial categorization and help to maintain separate AI models, their timing is concurrent, e.g., a location is associated to any user activity and state. But even within a context type, parallel processes exist, e.g. standing and shaking hands, and therefore often requiring individual AI recognition models and segmentation. One promising context recognition strategy is then to design one-vs-all or hierarchical pattern recognition models and report the top-K probable models.

However, the context modelling approach is application-dependent, given that in many cases the application requires a crisp single-state result.

From the perspective of data and information flow, a common context processing chain consists of the following stages (cf. Fig. 1):

1. **Data/signal acquisition:** Comprises various context sources, including physical sensors to measure phenomena properties, virtual sensors

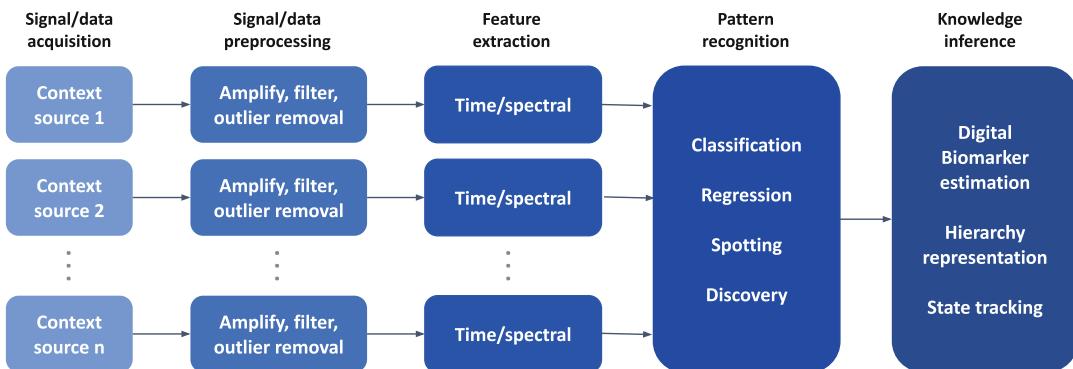


Fig. 1 Schematic of the context processing chain for wearable and implantable devices. The processing chain can be implemented with various AI components and for

all context types. Data input comes from device-integrated sensors and/or user interaction

that represent states of a system, as well as explicit and implicit interaction.

2. **Data preprocessing:** Depending on the context data source and data modality, different preprocessing is required. For physical sensors, the preprocessing may in part comprise signal amplification, signal filtering, and analog-to-digital conversion. For digital sources, algorithms include data-based filtering, outlier removal, and segmentation.
3. **Feature extraction:** Algorithms typically depend on the data modalities considered. Time-domain features include statistical signal representations, e.g. mean, variance, kurtosis, etc. Frequency-domain features are used to represent, e.g. electromyographic and acoustic signals. Other domains include wavelets and discrete spaces, e.g. for symbolic data representation. For an overview see, e.g., Figo et al. [20].
4. **Pattern recognition:** Algorithms range from supervised to unsupervised and typically target classification of context states, regression of a context variable, spotting of context events, or discovery of context states or events. Examples can be found in section “Selected Recognition Problems” below.
5. **Knowledge inference:** Whereas the pattern recognition stage is deriving discrete or continuous variables, the inference integrates variables into a (hierarchical) context representation, e.g. using topic models from text mining [21], or layered Hidden Markov Models (HMMs) [22]. By calibrating variables or further

estimation steps, digital biomarkers can be derived. Another task of this stage is variable tracking, i.e. to perform inference based on trends and states of recognition results.

Selected Recognition Problems

Context Pattern Spotting and Interpretation

A typical problem in medical context recognition is to identify events or specific physical or physiological situations in daily life using wearable and implantable systems. Context events may be basic state changes in variables, but mostly these are signal patterns with a temporal structure that may be similar to other, arbitrary data that surrounds the event patterns (referred to as *null class*) [23]. For example, food intake gestures show complex motion and orientation dynamics with variable duration in time. While intake gestures were modelled and recognized, e.g. using HMMs using wrist-worn devices, dealing with the null class in daily living data is highly challenging. To deal with the variability of null class patterns, a mined set of HMMs forming a network with the target gesture HMMs could be used [24]. The AI methods can be subsumed as pattern spotting, i.e. a combination of one-class recognition and efficient event search in continuous data.

As situation recognition tasks are typically intended to contextualize data and are followed

by further processing and interpretation steps, not all instances of the situation may be needed. In other words, it is frequently sufficient to attain a mediocre sensitivity, while maximizing the specificity or precision of the retrieval task, such that the further processing of biomarkers is minimally hampered by erroneous input. A typical situation recognition example is walking, indicating a condition of basic independence in or after rehabilitation [25], or physiological arousal [26]. For example, Derungs et al. used expert-designed rules to identify walking phases in hemiparetic patients, thus being dependent on the performance of features and thresholds only. The walking segments were then further interpreted regarding the patients' recovery. For this type of pattern spotting application, sensitivity must just be high enough to maintain a relevant biomarker update rate, i.e. event rate for spotting [24, 27].

Motion analysis is a long-standing topic in wearable computing systems and often tightly related to spotting specific events in data. In their 2020 systematic review, Weygers et al. [28] summarize a wide range of approaches to analyze joint kinematics and biomechanical modelling from inertial sensor units (IMUs). Their review considers preprocessing and information fusion aspects, besides dealing with drift, initialization, and calibration. They found that IMUs require further validation against gold standard methods, i.e. 3D gait analysis systems, and that the interpretation of IMU-based AI methods often builds on biomechanical assumptions and a-priori information. The authors expect that in the future individualized biomechanical joint models could be applied in populations with disturbed movement patterns in out-of-lab settings. Other recent approaches showed success in estimating gait mechanics, e.g. joint angle during motion, from trained Artificial Neural Networks (ANN) and IMU data [29].

Interpretation from Context Hierarchy

An interesting area of recognition problems are related to representing context as a multi-level concept, where top levels represent daily routines and subordinated layers deal with finer temporal

and spatial segmentation down to the level of atomic context, i.e. context that is not further subdivided and therefore must be derived from data sources. Pioneering work on context hierarchies have been developed in computer vision initially, e.g. [22]. Estimating cardiorespiratory fitness as an exemplary digital biomarker was investigated by Altini et al. [26], who considered two context layers (activity primitives, i.e. atomic context and a routine level) within a hierarchical Bayesian regression framework. Their approach demonstrated that the routine level adds essential information to contextualize heart rate for the regression with cardiorespiratory fitness. Similarly, a variety of approaches have built on hierarchical modelling, e.g. to describe daily routines or context composites with topic models [21, 30].

Routines occur in various contexts, where maintaining ordered action procedures is essential, e.g. clinical wards or the operating theatre. In their systematic review, Kolodzey et al. [31] summarized the technology and applications of wearables in the operating room, identifying use cases for telecommunication, safety, and information management, among others. For example, safety included augmented reality systems based on head-mounted displays and limb-worn devices for surgical navigation, and procedure tracking. They highlighted usability, inference robustness, data confidentiality, and runtime as challenges. Delivering in-time information and interaction options with head-mounted displays has been investigated, e.g. by Jalaliniya et al. [32], who focused on touchless, i.e. sterile, interaction, telepresence, and mobile patient record access. Input modalities included voice commands, head movements, and touch interaction. To detect and interpret the interaction forms correctly, context hierarchy information is essential. The authors noted that recognition speed is essential to leverage the interaction forms.

Flu Detection

A trend that started even before the COVID-19 pandemic is to estimate influenza-like symptoms from wearable devices. Radin et al. [33] showed in a population of ~47k individuals that estimates

of resting heart rate from activity trackers can improve the prediction of influenza-like illnesses reports with correlations ranging up to 0.97. Their approach build on linear regression and auto-regressive models to relate the heart rate data to reports. Using statistical summaries of tracker-based optical heartbeat, Li et al. [34] showed that an inflammation condition could be detected.

Major COVID-19 tracking approaches that build on the above insights mostly use expert knowledge modelling, primarily driven by the epidemiological characteristics of the studies, e.g. that no individual diagnostic data was available [5]. Mishra et al. [35] describe a study cohort with 5k participants using smartwatches to analyze physiological and activity data with signal alterations detectable in 80% of the COVID-19 infected cases. The physiological alternations appeared in 85% of the cases before symptoms became evident. The authors computed a HROS feature from heart rate and step count and applied an anomaly detection based on a Gaussian distribution assumption. Furthermore the online COSUM method was used to detect signal changes.

Situation Interpretation in Implants

Implantable computing systems are typically functioning as biosensors [36], organ support devices, e.g. autonomous pacemakers (APMs) or implantable cardioverter defibrillators (ICDs) [37], or ingestible monitoring or actuating system [38]. Implantable devices need to function with constraints on available system energy, potentially for several years, which limits computing resources. In addition, for actuating devices as APMs and ICDs, reproducibility of the intervention decisions are key for their risk assessment and medical device clearance. Simultaneously, several variables and patient conditions must be monitored for the devices to operate correctly. For example, inappropriate shocks of an ICD appear in about 15% of the patients and deteriorate quality of life for the patient [39].

ICD detection algorithms process electrocardiography (ECG) signals to determine the

timing of arterial and ventricular depolarization. A basic approach is to analyze the RR interval for detecting tachyarrhythmia, e.g. ventricular tachycardia (VT) [40]. However, further conditions exist, e.g. supraventricular tachycardia (SVT), which must be discriminated from VT to avoid inappropriate therapy. Overall, there are several conditions that can lead to VT or SVT. A typical implementation is thus a decision tree with three main branches [41]: (1) ventricular rate faster than atrial rate, denoting VT; (2) ventricular rate slower than atrial rate; and (3) both rates are equal. For conditions (2) and (3) various substates are considered to further subdivide the decision space. As the intracardiac signals vary with cardiac rhythms static sensing thresholds cannot be used in ICD devices. Instead, additional signal filter and estimation stages are implemented to adjust the thresholds dynamically.

Design and Construction

Wearable medical devices are designed to work in a biological environment that is challenging. Not only they have to be unobtrusive and autonomous, but they also have to fulfil ergonomics and comfort requirements. Among the new emerging technologies, AI can help designer in optimizing the wearable design and ensure a good quality product. According to the existing research work, AI has been applied to wearable design and construction to support developers at different stages of the process. The following sections provide an overview on how can the use of AI be beneficial to the development and fabrication. Along with the development of smarter wearable devices, design aspects such as the device geometry and material selection have gained more importance due to their impact on ergonomics and user acceptance. Although various mechanic and electronic Computer-Aided-Design (CAD) software are nowadays available to the designers, the integration of CAD with AI could further support the developers in the prototyping process and help solve fabrication issues ahead of time.

Topology Design and Optimization

When designing wearables with complex geometries, AI can check the presence of structural errors. For instance, in Additive Manufacturing (AM) process, CAD tools can perform quick design check and verify the device printability [42]. Lu et al. [43] proposed a Support Vector Machine-based (SVM) classifier that, based on features extracted from the design, i.e. structure complexity, as well as other fabrication parameters, i.e. printer model, could determine whether the design could be printed or not. The approach could be further developed by including other environmental indicators in the model, e.g. production time, cost, model dimensions etc. A Genetic Algorithm model could be used to analyze what factors influence the manufacturability and to which extent [44].

Despite AI could significantly reduce time and costs in designing complex geometries, to date its use has been limited and not fully investigated [45]. Yao et al. [46] proposed a hybrid machine-learning model for AM design feature recommendation. They proposed a hierarchical clustering for AM design features and target components and use a SVM classifier to determine the optimal design features. Although the work proved that AI can help inexperienced designers selecting the best AM design features, the proposed features had still to be tested through manual iterations and no automatic optimization process was implemented.

A more automatic approach to infer the optimal topology design was presented by Sosnovik et al. [47], who addressed the optimization problem by using a Convolutional Neural Network (CNN). Intermediate topologies from traditional topology optimization solvers based on Finite Element Method were used to train the CNN and predict after a few iterations the optimized structure. The method was later extended [48] to include 3D geometries. The trained model could predict optimized structures with 40% time reduction compared to traditional iterative topology optimisation approaches. With the use of Generative Adversarial Networks (GAN), it is also possible to predict the topology without using

iterative optimization. Given the design specifications, a trained GAN-model can generate possible multiple designs whose structure satisfy the set requirements [49]. However, training the model still relies on the results from standard topology optimization solvers.

In CAD-based apparel design, only basic fabric patterns can be automatically generated with AI. When more complex styles have to be developed, the standard approach still relies on decisions from experienced designers. Nevertheless, knowledge-based systems could ease developers in the design of a specific garment [50].

Wearable Personalization

In order to meet system performance, ergonomics, and comfort user needs, device personalisation should be taken into account during the design. A wrong fitting or misplacement on the body can invalidate measurements and impact vital signs monitoring [51]. Developers should adjust the wearable dimensions and shape and adapt them to end-user needs, yet the process requires precise anatomical measurements and thus can be tedious and time-consuming. A classic approach followed by manufacturers and researchers is to fabricate devices in different sizes and use geometry-adaptive materials to fit as many users as possible. However, in particular anatomical regions with complex morphology, such as ears, the optimal fitting could not be achieved with standardized device dimensions. Not only the optimal fitting but also the optimal sensor on-body position should be addressed. Personalized digital twins could help designers evaluating the sensor performance based on the on-body position, as proposed by Derungs and Amft. [52]. They analyzed the optimal position of inertial sensors to achieve best gait marker estimation in running athletes and hemiparetic patients.

To speed up the personalization, the wearable adjustment step could be optimized with the help of AI. Body measurements are automatically extracted with high accuracy by using photometric scanners that reconstruct the body 3D model from multi-perspective images or videos.

Afterwards, different models can be employed to improve the design. For instance, heuristic rules could be used to adjust the device parametric model based on the user's measures [53]. As a further development, ANN could be trained to output the best garment pattern design based on the input body model [50].

Prefabrication and Process Planning

After defining the wearable specifications and structure, other process fabrication parameters, including material choice need to be defined. Depending on the device requirements, some assumptions can be made earlier. When designing new, custom wearable prototypes, first experiments may imply several steps of trial-and-error to find the optimal parameters and material properties required for the application. The approach can be time-consuming and expensive, especially when conventional manufacturing techniques are employed for the construction. If an AM technique is used, simulation tools can help predicting parameters including mechanical robustness, production time, and amount of material needed. For instance, the 3D model can be sliced into planes orthogonal to z-axis to simulate the layer-by-layer deposition and evaluate the final quality of printed part based on fabrication settings i.e. layer height, nozzle diameter. In addition, the simulation could analyze the nozzle trajectory to determine whether the path could cause defects in the structure during the printing. The path optimization is an important aspect in the printing process, because it influences the fabrication time and final quality of the wearable [44]. In order to better support the parameter selection, a recent approach used in AI-enhanced AM is the use of Multi-Criteria Decision Making (MCDM) systems. In the MCDM approach, a set of possible parameters options including inter-factors relationships can be generated by existing databases or knowledge bases. The fabrication can be formulated as an optimization problem and solved based on user-defined requirements. For example, Wang et al. [54] proposed to use a hybrid MCDM system and rank possible solutions by employing a modified

technique for order preference by similarity to an ideal solution approach.

Besides supporting designers in the process and parameters selection, AI can be employed to predict the final mechanical material properties after the fabrication. Depending on AM parameters such as the infill pattern or density for 3D designs, properties like tensile and compressive strength may vary. Among the existing AI algorithms, the Multilayer Perceptron has been extensively applied to investigate macro-scale mechanical parameters. The analysis can be further extended to nano-scale properties when analyzing the surface conditions with classifiers such as SVM [45].

When employing smart garments as wearables, predicting the mechanical properties of fabric during manufacturing is fundamental to ensure quality devices. Specifically, it is important to evaluate upfront the seam efficiency. For example ANNs have been used to predict the textile performance during sewing [50].

Not only mechanical properties but also the fitting and comfort can be evaluated with AI. For instance, a Fuzzy-Logic system could predict fabric thermal comfort based on thermal and moisture sensations [55].

AI-Assisted Fabrication of Wearables

Although common fabrication processes are nowadays widely automatized, defects during the production could still occur, especially when using AM techniques. In particular, AM processing could cause defects including cracks, delamination, distortion, rough surface, and porosity. Most of the production failures originate from the layer deposition and selected process parameters. For example, interlayer adhesion or precise layer placement i.e. because of uncontrolled filament flow during the nozzle extrusion must be addressed. It is therefore beneficial a defect check during the construction.

AI can help developers by automating the in-process quality control. For this purpose, different sensing modalities have been tested by researchers to monitor AM. For instance, the

acoustic waves emitted from extruders i.e. in the Fused Deposition Modelling process during fabrication can be analyzed. The acoustic pattern could be examined with K-means clustering to determine if the printing was successful or anomalies occurred [45].

Among the existing monitoring approaches, optical-based techniques are mostly adopted. An infrared camera could be used to capture thermal images of the process and monitor the melt pools during Selective Laser Melting of metallic powders. Trained models including SVM or CNN could detect in thermal images defects like lack of fusion or keyhole porosity [45]. Alternatively, the wearable structure could be inspected by examining the surface condition or by employing CT images [45]. The visual inspection during the process could be also used in the fabrication of smart fabrics. ANNs have been extensively used to detect knitted fabric faults by comparing patterns with stored images of defects, e.g. see [50].

Validation

Despite the extensive research and large investments from healthcare stakeholders in medical wearable devices, a standardized framework for validation of wearable and implantable systems does not exist. The current, fragmented frameworks employed to validate AI-integrated wearable solutions cannot ensure product quality standards, i.e. effectiveness and low harm hazard [56].

AI System safety assessment

AI is commonly employed to support experts and users by automating iterative, analytical processes or augmenting data visualization. The more processes are automated, the more user decision or behavior relies on AI. Ideally, AI could augment any clinical decision-making process. If algorithms perform higher-level tasks, as proposing diagnosis, model error risk will increase as well as explainability due to task complexity [57]. It is therefore important to rigorously assess the

system accuracy and robustness during the device validation. When testing medical AI systems, retrospective cohort studies are widely used for the evaluation [57]. The study design usually consists of the following steps.

1. **Clinical question formulation:** First, the study objective is outlined. Typically, in the evaluation of AI systems the study aims to quantify how well the system can perform a certain task compared to a reference method. For instance, the accuracy in detecting dietary activity events with worn sensors could be the study objective [23]. As reference method, a standard medical practice is employed. Alternatively, the AI algorithm could be compared against opinions of human experts. For context-aware wearable and implantable systems, a variety of methods have been established to determine ground truth (see below).
2. **Ground truth:** After the study objective is defined, the approach to evaluate the AI model has to be designed. The AI output has to be compared to a known outcome, or ground truth, in order to determine if the algorithm response to a certain input is correct or not. Although ground truth information should be ideally perfect, missing data or errors can occur and decrease not only the AI model performance but also invalidate the performance assessment itself. It is therefore preferable to adopt objective measurements rather than human interpretations as ground truth. If human evaluation has to be used as ground truth, it is advisable to establish criteria to be followed from human raters that are as objective as possible. The rating criteria are necessary to avoid bias in the model. In addition, the ground truth should be based on the consensus opinion of more experts to maximize stability.
3. **Target population:** Subsequently, a population to be modelled by the AI algorithm has to be selected. The population includes patients, their clinical data as well as other indicators including sex, age, and other factors that are relevant to the study. For instance, if the wearable device aims to monitor digestion

and bowel motility [58], behavioral factors, e.g. dietary schedule, are relevant to study, because they influence the digestive process. The data from the target population are used to build and train the AI model. Hence, the dataset should be ideally large and uniformly distributed across all expected operational states of the AI model to avoid model bias or overfitting. Unfortunately, especially in medical AI systems, the availability of data is often limited and strategies need to be employed to increase the system robustness and generality. Data augmentation can be for example used when training deep ANN to overcome the dataset limited size. Combinations of data-driven and expert knowledge-based modelling can balance bias and variance in the system's performance, e.g. [59].

4. **Cohort:** Once the model is trained on a subset (train set) of the target population, the model performance is evaluated on another unseen subset (test set). Usually, at the beginning of the model design, the population is randomly divided into train set and test set. Since the train set is used to set the model parameters and the dataset size is generally limited, AI developers tend use a smaller subset of the original population as test set and employ most of the individuals in the training phase. Nevertheless, the cohort size should have an appropriate size to achieve a good generalizability of the results. Determining the proper cohort size is non-trivial and depends on the model task [57].
5. **Metrics:** A set of tools and methods are needed to assess the AI model performance. Typically, the model output tested on the cohort is compared against the ground truth. The most commonly used measurements are sensitivity (true positive rate) and specificity (true negative rate), which are often presented in visual form of a confusion matrix that for a binary classification resembles the truth table. Confusion matrices not only show correct predictions but also the algorithm mistakes are indicated in the form of false positive (negative cases predicted as positive) and false negative (positive cases predicted as negative). The algorithm performance should be evaluated

depending on the application goal. For instance, if the wearable is intended to be used for screening purposes, i.e. to detect COVID-19 symptoms [35], a highly sensitive algorithm should be preferred despite low specificity. The same metrics can be used also while training the model, in order to find the parameters yielding the best model performance. For various digital biomarker estimation tasks, continuous outputs are generated. To assess performance for continuous model outputs, error is measured based on distance metrics, e.g. signed or absolute difference, relative distance, etc.

Implantable System Testing

Given their elevated health risk, safety and certification procedures for implantable systems are more complex than those for wearables. The principal steps for implantable safety testing including the following, e.g. [60]:

1. **Concept checking:** During the conception of a new system design, experience of the developers is used to mentally assess the system's feasibility, including imagining its fabrication, its operation, and any potential issues.
2. **Computer simulation:** Ideally, the full system, including the AI functionality can be tested in co-simulations of the system and its application environment within the human body. The goal of the test is to assess whether all mechanisms operate as intended. Computer simulations are often limited by the fidelity of the models for the systems and the human body.
3. **Laboratory test:** Powering test on the benchtop are useful to assess fabrication quality and the matching of anticipated system properties and the AI models, e.g. whether the device's sensors produce data compatible with the AI model.
4. **Phantom test:** First operational and performance evaluation in a synthetic replica of the targeted environment within the human body. The testing produces data to re-validate the AI

models, but remains limited by the fidelity of the replica, e.g. soft silicon material to represent tissue.

5. **Ex-vivo test:** Where feasible, testing in actual biological tissue samples although under controlled laboratory conditions can offer insight on the complete system functionality. For example, interactions of complex tissue with sensors or actuators can be analyzed, communication to the body outside (when applicable) can be tested, etc.
6. **Cadaver/animal model test:** The closest testing to in-vivo human implantation is to assess feasibility and performance under the actual condition of a complete target environment. Algorithm updates are still feasible in this step. Human cadavers are donated bodies for testing medical procedures. Animal models provide similar environments to the targeted one within the human body. For example, certain pigs types have a similar gastrointestinal tract to humans and thus could be used for final stage testing of ingestible devices.
7. **In-human test:** After regulatory review and meeting all ethical and safety standards, implantable devices are finally tested in humans volunteers during a clinical trial. The test is providing additional qualification data for medical device risk assessment and certification by authorities.

Self-Checking AI Systems

Wearable device robustness and accuracy could degrade over time due to sensors failures or device displacement. An advantage of embedding AI on wearable devices is the possibility to not only calibrate the sensor based on the user, but also to adapt the measurements in case of device displacements. The detection and compensation of anomalies in the sensor readings can therefore ensure that degradation of inference functions are avoided. For instance, integrating information from multiple data sources within a decision tree [61] could help identifying sensor errors. While the accidental device displacement could decrease, i.e. accelerometer reliability, rotation

data from gyroscopes could compensate reading errors and avoid false detections. Furthermore, unsupervised methods can be employed to adapt for motion-induced drifts, without requiring information redundancy from multiple sensors. For example, if the data distribution shifts over time, an Expectation-Maximization-based algorithm can be applied to recover the original distribution, even online during system deployment [62].

Usability

Since most of the medical wearable devices are designed to operate in a long-term setting, it is important to ensure user acceptance and engagement by evaluating and testing the wearable usability. Firstly, the usability could be reviewed according to human experts. For instance, Jakob Nielsen in 1994 defined ten heuristic rules to help developers in the usability evaluation of interfaces. Additionally, the wearable usability could be tested by asking volunteers to use the wearable within a clinical trial and make them compiling questionnaires afterwards. A commonly used questionnaire is the System Usability Scale, proposed by Brooke in 1996 [63]. The form includes eight questions rating the system usability and two additional questions rating the system learnability, and it proved helpful to evaluate human computer interaction-based systems, i.e. mobile apps. However, the mentioned methods were design to mainly tests user interfaces. A more general procedure to evaluate wearable and implantable systems usability, including less interactive systems, is still missing and should be further developed.

References

1. Amft O. How wearable computing is shaping digital health. IEEE Pervasive Comput. 2018;17(1):92–8.
2. Banaee H, Ahmed MU, Loutfi A. Data mining for wearable sensors in health monitoring systems: a review of recent trends and challenges. Sensors. 2013;13(12):17472–500.
3. Sankar PL, Parker LS. The precision medicine initiative's all of us research program: an agenda for research on its ethical, legal, and social issues. Genet Med. 2017;19(7):743–50.

4. Byambasuren O, Sanders S, Beller E, Glasziou P. Prescribable mHealth apps identified from an overview of systematic reviews. *NPJ Digital Med.* 2018;1(1):12.
5. Amft O, Lopera Gonzalez LI, Lukowicz P, Bian S, Burggraf P. Wearables to fight COVID-19: from symptom tracking to contact tracing. *IEEE Pervasive Comput.* 2020;19(4):53–60.
6. Qi J, Yang P, Min G, Amft O, Dong F, Xu L. Advanced internet of things for personalised healthcare systems: a survey. *Pervasive Mobile Comput.* 2017;41:132–49.
7. Amft O, Laerhoven KV. What will we wear after smartphones? *IEEE Pervasive Comput.* 2017;16(4):80–5.
8. Reiss A, Amft O. Design challenges of real wearable computers. In: Barfield W, editor. *Fundamentals of wearable computers and augmented reality*. CRC Press; 2015. p. 583–618.
9. Summers GD, assignee. Implantable bio-data monitoring method and apparatus. US3672352A. 1972.
10. Kawsar F, Min C, Mathur A, Montanari A. Earables for personal-scale behavior analytics. *IEEE Pervasive Comput.* 2018;17(3):83–9.
11. Park J, Kim J, Kim SY, Cheong WH, Jang J, Park YG, et al. Soft, smart contact lenses with integrations of wireless circuits, glucose sensors, and displays. *Sci Adv.* 2018;4(1):eaap9841.
12. Chen C, Zhao XL, Li ZH, Zhu ZG, Qian SH, Flewitt AJ. Current and emerging technology for continuous glucose monitoring. *Sensors.* 2017;17(1):182.
13. Bailey CJ, Gavin JR. Flash continuous glucose monitoring: a summary review of recent real-world evidence. *Clin Diabetes.* 2020.
14. Schilit B, Adams N, Want R. Context-Aware computing applications. In: WMCSA 1994: IEEE Workshop on Mobile Computing Systems and Applications; 1994. p. 89–101.
15. Krumm J. Ubiquitous computing fundamentals. 1st ed. Chapman & Hall/CRC; 2009.
16. Chen G, Kotz D. A survey of context-aware mobile computing research. Dartmouth College: Hanover; 2000.
17. Dey AK, Abowd GD, Salber D. A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Hum Comput Interact.* 2001;16(2):97–166.
18. Yürür Ö, Liu CH, Sheng Z, Leung VCM, Moreno W, Leung KK. Context-awareness for mobile sensing: a survey and future directions. *IEEE Commun Surv Tutorials.* Firstquarter. 2016;18(1):68–93.
19. Lukowicz P, Junker H, Stäger M, von Büren T, Tröster G. WearNET: a distributed multi-sensor system for context aware wearables. In: Goos G, Hartmanis J, van Leeuwen J, editors. *Ubicomp 2002: Proceedings of the 4th International Conference on Ubiquitous Computing*. vol. 2498. Berlin/Heidelberg: Springer; September–October 2002. p. 361–370.
20. Figo D, Diniz PC, Ferreira DR, Cardoso JMP. Pre-processing techniques for context recognition from accelerometer data. *Pers Ubiquit Comput.* 2010;14(7):645–62.
21. Seiter J, Amft O, Rossi M, Tröster G. Discovery of activity composites using topic models: an analysis of unsupervised methods. *Pervasive Mobile Comput.* 2014;15:215–27.
22. Oliver N, Garg A, Horvitz E. Layered representations for learning and inferring office activity from multiple sensory channels. *Comput Vis Image Underst.* 2004;96(2):163–80.
23. Amft O, Tröster G. Recognition of dietary activity events using on-body sensors. *Artif Intell Med.* 2008;42(2):121–36.
24. Schiboni G, Amft O. Sparse natural gesture spotting in free living to monitor drinking with wrist-worn inertial sensors. In: *Proceedings of the 2018 ACM International Symposium on Wearable Computers*. ISWC '18. New York: ACM; 2018. p. 140–147.
25. Derungs A, Schuster-Amft C, Amft O. Longitudinal walking analysis in Hemiparetic patients using wearable motion sensors: is there convergence between body sides? *Front Bioeng Biotechnol.* 2018;6.
26. Altini M, Casale P, Penders J, Amft O. Cardiorespiratory fitness estimation in free-living using wearable sensors. *Artif Intell Med.* 2016;68:37–46.
27. Amft O. Adaptive activity spotting based on event rates. In: *SUTC 2010: Proceedings of the IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing*. IEEE; 2010. p. 169–176.
28. Weygers I, Kok M, Konings M, Hallez H, De Vroey H, Claeys K. Inertial sensor-based lower limb joint kinematics: a methodological systematic review. *Sensors.* 2020;20(3):673.
29. Mundt M, Koeppe A, David S, Witter T, Bamer F, Potthast W, et al. Estimation of gait mechanics based on simulated and measured IMU data using an artificial neural network. *Front Bioeng Biotechnol.* 2020;8.
30. Huynh T, Fritz M, Schiele B. Discovery of activity patterns using topic models. In: *UbiComp 2008: Proceedings of the 10th International Conference on Ubiquitous Computing*. vol. 344 of *ACM International Conference Proceeding Series*. Seoul: ACM; 2008. p. 10–19.
31. Kolodzey L, Grantcharov PD, Rivas H, Schijven MP, Grantcharov TP. Wearable technology in the operating room: a systematic review. *BMJ Innov.* 2016;3(1):55–63.
32. Jalaliniya S, Pederson T, Mardanbegi D. A wearable Personal Assistant for surgeons – design, evaluation, and future prospects. *EAI Endorsed Trans Pervasive Health Technol.* 2017;3(12):153066.
33. Radin JM, Wineinger NE, Topol EJ, Steinhubl SR. Harnessing wearable device data to improve state-level real-time surveillance of influenza-like illness in the USA: a population-based study. *Lancet Digital Health.* 2020;2(2):e85–93.
34. Li X, Dunn J, Salins D, Zhou G, Zhou W, Rose SMSF, et al. Digital health: tracking Physiomes and activity

- using wearable biosensors reveals useful health-related information. *PLoS Biol.* 2017;15(1):e2001402.
35. Mishra T, Wang M, Metwally AA, Bogu GK, Brooks AW, Bahmani A, et al. Early detection of COVID-19 using a smartwatch. 2020;medRxiv <https://doi.org/10.1101/2020.07.06.20147512>.
36. Rodrigues D, Barbosa AI, Rebelo R, Kwon IK, Reis RL, Correlo VM. Skin-integrated wearable systems and implantable biosensors: a comprehensive review. *Biosensors.* 2020;10(7):79.
37. Ammannaya GKK. Implantable cardioverter defibrillators – the past, present and future. *Arch Med Sci Atherosclerotic Dis.* 2020;5:e163–70.
38. Kiourti A, Psathas KA, Nikita KS. Implantable and ingestible medical devices with wireless telemetry functionalities: a review of current status and challenges. *Bioelectromagnetics.* 2014;35(1):1–15.
39. Seyis S, Kurmus O, Turhan H. Clinical utility of an implantable cardioverter-defibrillator lead with a floating atrial sensing dipole: a single-center experience. *Arch Clin Biomed Res.* 2017;1(6):289–300.
40. Swerdlow CD. Supraventricular tachycardia-ventricular tachycardia discrimination algorithms in implantable cardioverter defibrillators: state-of-the-art review. *J Cardiovasc Electrophysiol.* 2001;12(5):606–12.
41. Brüggemann T, Dahlke D, Chebbo A, Neumann I. Tachycardia detection in modern implantable cardioverter-defibrillators. *Herzschriftmacherther Elektrophysiol.* 2016;27(3):171–85.
42. Tansaz S, Baronetto A, Zhang R, Derungs A, Amft O. Printing wearable devices in 2D and 3D: an overview on mechanical and electronic digital co-design. *IEEE Pervasive Comput.* 2019;18(4):38–50.
43. Lu T. Towards a fully automated 3D printability checker. In: 2016 IEEE International Conference On Industrial Technology (ICIT). IEEE; 2016. p. 922–27.
44. Yang J, Chen Y, Huang W, Li Y. Survey on artificial intelligence for additive manufacturing. In: 2017 23rd International Conference on Automation and Computing (ICAC). IEEE; 2017. p. 1–6.
45. Wang C, Tan XP, Tor SB, Lim CS. Machine learning in additive manufacturing: state-of-the-art and perspectives. *Addit Manuf.* 2020;36:101538.
46. Yao X, Moon SK, Bi G. A hybrid machine learning approach for additive manufacturing design feature recommendation. *Rapid Prototyp J.* 2017;23(6):983–97.
47. Sosnovik I, Oseledets I. Neural networks for topology optimization. *Russ J Numer Anal Math Model.* 2019;34(4):215–23.
48. Banga S, Gehani H, Bhilare S, Patel S, Kara L. 3D topology optimization using convolutional neural networks. arXiv preprint arXiv:180807440. 2018.
49. Rawat S, Shen MHH. A novel topology design approach using an integrated deep learning network architecture. arXiv preprint arXiv:180802334. 2018.
50. Nayak R, Padhye R. Artificial intelligence and its application in the apparel industry. In: *Automation in Garment Manufacturing.* Elsevier; 2018. p. 109–38.
51. Harms H, Amft O, Tröster G. Does loose fitting matter? Predicting sensor performance in smart garments. In: *Bodynets 2012: Proceedings of the International Conference on Body Area Networks.* ACM; 2012. p. 1–4.
52. Derungs A, Amft O. Estimating wearable motion sensor performance from personal biomechanical models and sensor data synthesis. *Nat Sci Rep.* 2020;10(11450). <Https://rdcu.be/b5ynj>
53. Wahl F, Zhang R, Freund M, Amft O. Personalizing 3D-printed smart eyeglasses to augment daily life. *IEEE Comput.* 2017;50(2):26–35.
54. Wang Y, Zhong RY, Xu X. A decision support system for additive manufacturing process selection using a hybrid multiple criteria decision-making method. *Rapid Prototyp J.* 2018;24(9):1544–53.
55. Wang Z, Li Y, Wo Wong AS. Simulation of clothing thermal comfort with fuzzy logic. In: *Elsevier Ergonomics Book Series.* vol. 3. Elsevier; 2005. p. 467–71.
56. Mathews SC, McShea MJ, Hanley CL, Ravitz A, Labrique AB, Cohen AB. Digital health: a path to validation. *NPJ Digital Med.* 2019;2(1):1–9.
57. Oakden-Rayner L, Palmer LJ. Artificial intelligence in medicine: validation and study design. In: Ranschaert ER, Morozov S, Algra PR, editors. *Artificial intelligence in medical imaging: opportunities, applications and risks.* Cham: Springer International Publishing; 2019. p. 83–104.
58. Baronetto A, Graf LS, Fischer S, Neurath MF, Amft O. GastroDigitalShirt: a smart shirt for digestion acoustics monitoring. In: *ISWC '20: Proceedings of the 2020 International Symposium on Wearable Computers.* Virtual Conference: ACM; 2020. p. 17–21.
59. Wahl F, Amft O. Data and expert models for sleep timing and chronotype estimation from Smartphone context data and simulations. *Proc ACM Interact Mob Wearable Ubiquitous Technol.* 2018;2(3):139:1–139:28.
60. Pietro Valdastri. Sonopill: Test Environments [Internet]. Test Environments. 2018 [cited 2019 Oct 31]. Available from: Https://www.gla.ac.uk/research/az/sonopill/blog/headline_569250_en.html
61. Reiss A, Cheng J, Amft O. Hierarchical motion artefact compensation in smart garments. In: *Proceedings of 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing Workshops.* ACM; 2014. p. 243–48.
62. Bayati H, J del R Millan, Chavarriaga R. Unsupervised adaptation to on-body sensor displacement in acceleration-based activity recognition. In: *Proceedings of the 15th Annual International Symposium on Wearable Computers.* IEEE press; 2011. p. 71–8.
63. Brooke J. SUS-A quick and dirty usability scale. *Usability Eval Ind.* 1996;189(194):4–7.



Machine Learning and Electronic Noses for Medical Diagnostics

86

Wojciech Wojnowski and Kaja Kalinowska

Contents

A Brief Introduction to the Electronic Nose Technique	1204
An Overview of Electronic Nose Application in Medical Diagnostics	1206
Machine Learning in Enhancing the Diagnostic Capability of Electronic Noses	1210
Denoising Algorithms	1211
Sensor Drift Compensation	1211
Conclusions	1213
References	1214

Abstract

The need for noninvasive, easy-to-use, and inexpensive methods for point-of-care diagnostics of a variety of ailments motivates researchers to develop methods for analyzing complex biological samples, in particular human breath, that could aid in screening and early diagnosis. There are hopes that electronic noses, that is, devices based on arrays of semi-selective or nonselective chemical sensors, can fill this niche. Electronic olfaction uses data processing and machine learning to build classification models based on the responses of

several sensors in the form of multivariate datasets in order to discriminate between disease and healthy control based on a unique fingerprint. However, the introduction of this technique in clinical settings is limited by methodological issues which can, to some extent, be remedied using artificial intelligence. In this chapter, we provide a brief introduction to the electronic nose technique and outline its applications in medical diagnostics. We also discuss the ways in which data processing and machine learning techniques can be used to facilitate the use of electronic olfaction in the detection of disease.

Keywords

Electronic noses · Electronic olfaction · Medical diagnostics · Breath analysis · Chemical sensors · Sensor drift · Pattern recognition · Machine learning

W. Wojnowski (✉) · K. Kalinowska
Department of Analytical Chemistry, Gdańsk University of
Technology, Gdańsk, Poland
e-mail: wojciech.wojnowski@pg.edu.pl;
kaja.kalinowska@pg.edu.pl

A Brief Introduction to the Electronic Nose Technique

The term “electronic nose” usually refers to a technique in which a device equipped with an array of different chemical gas sensors is used to discriminate between complex gaseous mixtures based on their unique fingerprint, i.e., the characteristic pattern in the response of the sensor array. The concept in its present form was introduced by Persaud and Dodd in the early 1980s [1], and the term “electronic nose” is commonly used in the literature, although some authors point out that it suggests too close a parallel with the mammalian olfactory sense and that “multisensorial system” or “gas sensor array” are more suitable names [2, 3]. One could argue, however, that the former is more general and could refer to, for example, electronic tongues [4] or other techniques not necessarily used to analyze gaseous samples, while the latter describes the detector element of the conventional electronic noses. Still, the comparison with the operation of the human sense of smell might be useful to illustrate the principle of operation of electronic olfaction. In the mammalian olfactory system, after sample acquisition, either passive (odorants reaching the human nose through diffusion) or active, by inhalation, the volatile substances come in contact with specialized, selective chemoreceptors located in the nasal cavity, in the olfactory epithelium. The axons of these sensory neurons then project to the olfactory bulb, where they produce spatial patterns of activity which overlap to a certain extent, but vary for different odors [5]. Similarly, in electronic noses, the gaseous sample is collected either passively or, more commonly, actively by means of a small vacuum pump and directed into a sensor chamber in which the analytes come in contact with an array of chemical sensors (see Fig. 1). The response signals of these sensors are in turn digitized, and the characteristic response patterns of the entire array are analyzed using data processing and pattern recognition algorithms [6]. Here, however, a major difference between biological and chemical olfaction has to be indicated: while the biological chemoreceptors are selective [7], a vast majority of chemical

sensors used in the e-nose sensor matrices are either nonselective or partially selective to particular chemical compounds. Furthermore, chemical sensors can respond to chemicals which are not perceivable by the mammalian sense of smell or to odorants below the human odor threshold.

The main advantage of using electronic olfaction is its versatility which stems from the ability to tailor the sensor array to a particular sample or application drawing from a wide range of commercially available and relatively inexpensive chemical sensors. As these devices usually do not require sophisticated peripherals and can themselves be relatively uncomplicated, portable, and easy to operate without extensive training, they can be a valid, low-cost alternative to various instrumental analytical techniques used in applications requiring the analysis of gaseous mixtures, e.g., in food authentication and quality monitoring, environmental comfort and safety, or medical diagnostics. Furthermore, they can be geared toward nondestructive and remote sensing. This is particularly relevant in the context of equitable chemistry [8], as it could greatly increase the accessibility to analytical tests in everyday life, especially since the computational power required to process the obtained data, and potentially also the user interface, is available in the form of ubiquitous smartphones.

Since electronic noses are usually used to analyze relatively complex gaseous matrices and since the chemical sensors which comprise their detector system are, on the most part, only partially selective, it is difficult to categorize the obtained analytical information as qualitative or quantitative. Instead, the response signals of the sensors are treated holistically as a characteristic fingerprint or rather “smellprint” of the sample. This fingerprint can then be compared with an existing reference library either to classify the sample or to perform a regression task based on a target reference value. This is achieved through AI and more specifically through pattern recognition algorithms and machine learning (ML).

In practical terms, the raw signals of the sensor array are passed through an analogue-digital converter and registered using a computer or an equivalent device. After a series of measurements,

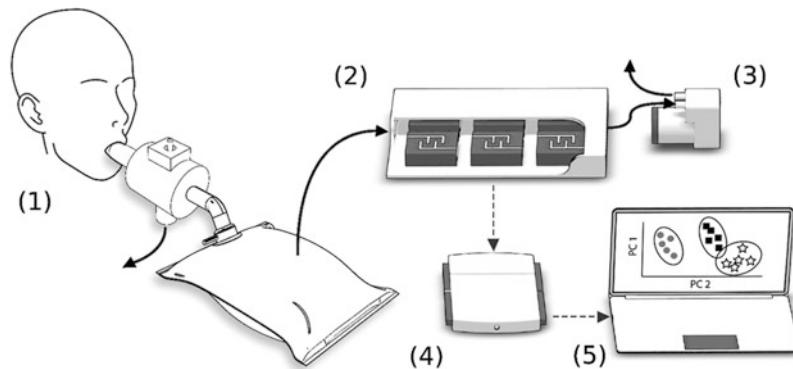


Fig. 1 A schematic representation of an electronic nose used for analyzing a breath sample: (1) sampling setup consisting of a device that captures only the end-tidal breath fraction based on the measurement of CO₂ concentration and a Tedlar bag, (2) an array of six different

chemical sensors with (3) active sampling using a vacuum pump, (4) analogue-digital converter, (5) data pre-processing and feature extraction. (Modified from [10], Copyright 2018, with permission from Elsevier)

a dataset is obtained, with signals from each sensor constituting a separate variable and values obtained in each run constituting separate objects. In order to process this multidimensional data, it is usually necessary to use multivariate statistical analysis. The aim of this process is to develop a model trained using machine learning algorithms that will be able to either identify an unknown sample as belonging to a particular discrete category based on its unique “fingerprint” or to predict a continuous numerical value corresponding to a result of a reference method. The entire procedure could be separated into distinct stages, namely, (i) preprocessing, (ii) feature extraction, (iii) development and validation of a machine learning model, and finally (iv) classification [9].

Preprocessing usually involves filtering the signal of the sensors to remove noise, temperature, and humidity corrections and defining the response. The raw response of an electronic nose consists of a number of time series equal to the number of sensors in the array. These time series are usually characterized by a transient state during which the signal changes as the sensors are exposed to the analytes, followed by a steady state during which the signal becomes nearly constant. How soon this happens depends on multiple factors, of which perhaps the most important one is the characteristic of the sensors themselves. In most electronic noses, analytically useful signal

is obtained in under 5 min and often in approximately 1 min [11]. The most straightforward procedure is to subtract the baseline signal recorded before the sensors were exposed to the analytes from the response in the steady state after sampling, but other approaches are also utilized [12]. Next, the data is normalized in order to increase the performance of further data processing and ML stages which might be sensitive to noncomparable variables. This is commonly achieved by centering each feature and then scaling the values by standard deviation, in the range -1:1 or 0:1. In cases in which the ML model will utilize an algorithm geared toward classification, continuous target variables assigned to the dataset might also be discretized at that stage.

After the data is converted into the feature space, preferably with several objects (discrete measurements), the input features might be transformed into new, salient features [13]. This is often done in order to reduce the dimensionality of the dataset, i.e., to lower the ratio of the number of features to the number of objects, which is particularly important in the case of e-noses which are not based on an array of chemical sensors and instead extract the features, e.g., from a mass spectrum or ion mobility spectrum, and where the number of features might easily exceed the number of objects leading to overfitting of the ML model [14]. In such a case, the

model shall reflect the idiosyncrasies of the training data and underperform in real-life applications [15]. The most commonly employed technique for dimensionality reduction is principal component analysis (PCA). It creates a new coordinate system in a way that maximizes the variance along with the first and subsequent new coordinates (component), each orthogonal to the previous one. The result is a new set of variables in which a majority of the variance of the original dataset can be explained by a couple of the new features. Other techniques for dimensionality reduction include linear discriminant analysis, fast Fourier transform, discrete wavelet transform, clustering in feature space, etc. [16]. In cases in which the sensor array of the electronic nose was not tailored to the specific application and so it likely contains sensors that do not contribute to the analysis of the particular matrix, the most relevant features with relation to a target reference value can also be selected using analysis of variance (ANOVA).

The last and perhaps most important stage of processing the data obtained using electronic noses is the use of pattern recognition algorithms. They can be broadly categorized based on their use for either classification or regression task, with an overlap in methods used in both applications. It is important to note that since the response of the sensor matrix is difficult to describe qualitatively and quantitatively, especially in the case of complex matrices such as human volatilome, the supervised pattern recognition algorithms require a target variable obtained using a reference method for training. This variable can be either categorical or continuous – hence, the distinction between classification and regression tasks. In the case of medical diagnostics, this target variable could be as straightforward as diagnosis (patient/control) or a parameter determined using a more time-consuming method, such as the concentration of short-chain fatty acids in feces for the purpose of diagnosing inflammatory bowel disorders [10]. The algorithms which are most commonly used in ML in electronic olfaction include artificial neural networks (ANN), support vector machines (SVM), random forest (RF), k-nearest neighbors (k-NN) classifiers, and Bayesian networks. The trained models are then validated,

which is best done using a dataset that was obtained in a separate experiment [17]. Other approaches include drawing separate training and testing datasets from the same dataset, leave-one-out validation, and stratified cross-validation.

While pattern recognition ML methods for classification and regression tasks are the main and crucial application of artificial intelligence in electronic olfaction, they are not specific for its application in medical diagnostics. For a more comprehensive review of this topic, the reader is referred to [12, 18, 19]. In the following sections, we focus instead on the use of AI in mitigating the issues which hinder the use of electronic noses in medical diagnostics and which mostly stem from the limitations of gas sensors and from the complexity of the matrix, i.e., sensor drift, effect of humidity, and background volatiles in point-of-care testing.

An Overview of Electronic Nose Application in Medical Diagnostics

The use of the human sense of smell to detect indicators of disease was mentioned as early as 400 BCE, and the olfactory diagnosis was commonly used in traditional Chinese medicine [20]. This is unsurprising since various disorders might cause changes in the human volatilome which can then be used as diagnostic indicators. The successful use of trained dogs to recognize urine of cancer patients [21] has spurred interest in the development of diagnostic methods based on the analysis of fingerprints of breath and the headspace of other biological samples. While it is important to reiterate that electronic noses are noses by name only and the sensor matrices react to chemical compounds which do not necessarily contain odorants, they are nonetheless well-suited for discriminating between complex mixtures of volatiles. Ideally, a disease could be diagnosed based on the determination of a particular chemical compound which is a characteristic indicator of the particular ailment. Unfortunately, this is almost never the case, and the products of the metabolic, oxidative, and inflammatory processes which underlie various disorders are linked

with several interrelated biochemical reactions which cannot be investigated at the level of single molecules [22]. This is why data processing and pattern recognition techniques are used to identify subtle differences between samples from patients and from the control group.

Gold standards in the field of various ailment detection are usually methodologies that are not only time-consuming and expensive but also cumbersome for the patients. The explicit promise of electronic olfaction in medical diagnostics is the introduction of ubiquitous, relatively inexpensive, and easy-to-operate screening tests which would provide the result within minutes and would not require invasive or otherwise troublesome sampling (e.g., collecting stool samples, which might be taboo for some patients [23]). Thus, the emphasis is increasingly being placed on creating more sophisticated methodologies that would be less troublesome for the afflicted and could, in the future, replace methods such as colonoscopies, tissue biopsies, or multiple blood tests.

In recent years, numerous studies concerning disease detection based on the composition of the exhaled breath or urine, feces, or saliva headspace were carried out with the use of electronic noses. These devices seem well-suited for near real-time monitoring (a single sampling-purging cycle can be realized in less than 2 min) which opens new possibilities in point-of-care diagnostics and could also greatly improve the access to basic testing at first contact medical facilities and field screening tests [13].

E-noses have been successfully applied in multiple studies concerning numerous diseases, especially but not limited to those related to the respiratory system. They were used to distinguish between healthy controls and patients with, for example, asthma, COPD, and lung cancer [24–27], as well as to monitor the severity of the illness or detect the presence of infections during their exacerbation [28, 29]. Moreover, in several cases with the use of electronic nose, it was possible to differentiate between diseases that might manifest themselves in similar ways, such as allergic rhinitis and asthma [30] or COPD and asthma [31], which seems to further corroborate the assumption that electronic olfaction might find

application in routine screening, preliminary diagnostics, or monitoring whether the disease aggravates or alleviates.

While in some studies e-noses were used to analyze the headspace of feces, urine, or wounds, disease diagnostics is, most commonly, performed based on the analysis of exhaled air. Breath samples are collected in inert bags or delivered through a mouthpiece directly to the electronic nose. However, in certain cases, the electronic nose is designed to have a module that allows unmediated breath sampling, as is the case with, for example, SpiroNose (Breathomix, Leiden), which is a coupling of e-nose and spirometer [32]. By far the most popular technique for reducing the dimensionality of the dataset prior to the development of the ML model is principal component analysis (PCA). However, the application of analysis of variance (ANOVA) or covariance (ANCOVA) can be useful in experiments in which the difference are the main focus as is the case with research concerning differentiation between sick patients and healthy controls [33]. In the majority of reviewed studies, receiver operating characteristic (ROC) curve is utilized to assess the performance of the classifier – something that is more commonly seen in medical research than in other areas of e-nose applications. It would seem that artificial neural networks (ANN) in their various guises are becoming increasingly popular compared to other ML algorithms. It is mostly due to its appropriateness for situations in which additional data points are added over time and retraining is necessary [34]. This characteristic of neural networks would be a major advantage in the instance of possible future real-life application of electronic noses for routine screening and disease diagnostic purposes. Selected applications of electronic olfaction in medical diagnostics since 2015 are listed in Table 1. It supplements a previous, similar listing with more recent studies [13]. Detailed reviews of the various applications of electronic olfaction in medical diagnostics can be found in the works of Wilson and others [20, 35–40].

Despite these advances and the seeming suitability of electronic olfaction for rapid, noninvasive diagnostic tests based primarily on breath

Table 1 Selected applications of electronic noses in medical diagnostics

Diagnostic application	Matrix	Sampling method	Data analysis	Ref
Airway inflammation	Breath	Tedlar bag	PCA, k-nn	[42]
Allergic rhinitis	Breath	Tedlar bag	PCA, CDA, ROC curve	[43]
Amyotrophic lateral sclerosis	Breath	Inert bag connected to the e-nose	PCA, ANOVA, CDA, ROC curve	[44]
Aspergillus fumigatus	Breath	Tedlar bag	PCA, ROC curve	[45]
Asthma	Breath	Mouthpiece connected to the e-nose	PCA, logistic regression, ROC curve	[46]
Asthma	Breath	Tedlar bag	PCA, PLSDA	[29]
Asthma	Breath	Tedlar bag	PCA, ROC curve	[24]
Asthma	Breath	Mouthpiece connected to the e-nose	ANN	[47]
Asthma	Breath	Tedlar bag	PCA, ANCOVA	[48]
Asthma	Breath	Mouthpiece connected to the e-nose	PCA, ANOVA	[49]
Asthma	Breath	Tedlar bag	PCA, CDA, ROC curve	[30]
Chronic kidney disease	Breath	Tedlar bag	PCA, SVM, HCA, PLS-regression	[50]
Colorectal cancer	Urine	Urine headspace	LDA	[51]
COPD	Breath	Tedlar bag	PCA, ANCOVA	[26]
COPD	Breath	Mouthpiece connected to the e-nose	ANOVA, PLSDA	[52]
COPD	Breath	Mouthpiece connected to the e-nose	PCA, ANOVA	[31]
COVID-19	Breath	Mouthpiece connected to the e-nose	ANN	[53]
COVID-19	Breath	Mouthpiece connected to the e-nose	DFA, ROC curves	[54]
Cystic fibrosis	Breath	Mouthpiece connected to the e-nose	ANN	[47]
Diabetes	Breath	Tedlar bag	PCA, k-nn	[42]
Diabetes mellitus	Breath	Tedlar bag	PCA, SVMs, HCA, PLS-regression	[50]
Gastric cancer	Breath	Mouthpiece connected to the e-nose	ANN	[55]
Head and neck, bladder, and colon carcinomas	Breath	Mouthpiece connected to the e-nose	ANN	[56]
Head, lung, and neck carcinoma	Breath	Mouthpiece connected to the e-nose	ANN	[57]
Infection during COPD exacerbation	Breath	Tedlar bag	LDA, SLR	[28]
Infections in acute COPD exacerbations	Breath	Mouthpiece connected to the e-nose	ANN	[58]
Inflammatory asthma	Breath	Bag	PCA, ROC curves	[59]
Inflammatory bowel disease	Breath	Bio-VOC connected to the e-nose	SVMs, RF, SLR, ANN	[60]
Interstitial lung disease	Breath	Mouthpiece connected to the e-nose	PLSDA, ROC curve	[61]
Locoregional recurrent head and neck squamous cell carcinoma	Breath	Mouthpiece connected to the e-nose	ANN	[62]

(continued)

Table 1 (continued)

Diagnostic application	Matrix	Sampling method	Data analysis	Ref
Lung cancer	Breath	Tedlar bag	LDA, SVM, ROC curve	[63]
Lung cancer	Breath	Bag	GRU-AE-MSEP framework	[64]
Lung cancer	Breath	Tedlar bag and mouthpiece connected to the e-nose	k-nn, RF, LDAm SVM	[25]
Lung cancer	Breath	Tedlar bag	PLSDA, ROC curves	[65]
Lung cancer	Breath	Bag	SVMs	[66]
Lung cancer	Breath	Mouthpiece connected to the e-nose	ANN	[67]
Lung cancer	Breath	Bag	DFA, K-fold cross-validation, CART	[68]
Malignant pleural mesothelioma	Breath	Tedlar bag	PCA, ROC curve	[69]
Necrotizing enterocolitis	Feces	Feces' headspace	PCA, ROC curve	[70]
Obstructive sleep apnea	Breath	Tedlar bag	k-nn, CDA, ROC curves	[71]
Esophageal adenocarcinoma	Breath	Mouthpiece connected to the e-nose	Cross-validated prediction model	[72]
Overlap syndrome [obstructive sleep apnea with COPD]	Breath	Inert bag	PCA, CDA, ANOVA	[73]
Parkinson's disease	Breath	Mylar bag	DFA	[74]
Persistent asthma	Exhaled breath condensate	Condensate's headspace	PCA, ROC curve	[75]
Pneumoconiosis	Breath	FlexFoil Plus bag	LDA, SVMs	[76]
<i>Pseudomonas aeruginosa</i> and airway bacterial colonization in bronchiectasis	Breath	Tedlar bag	PCA, ANOVA, ROC curve	[77]
Psoriatic arthritis	Breath	Tedlar bag	PCA, DA, AUC, ROC curve	[78]
Renal failure	Breath	Tedlar bag	PCA, k-nn	[42]
Response to small-cell lung cancer treatment prediction	Breath	Mouthpiece connected to the e-nose	ROC curve	[79]
Rheumatoid arthritis	Breath	Tedlar bag	PCA, DA, AUC, ROC curve	[78]
Seasonal allergic rhinitis	Breath	Mouthpiece connected to the e-nose	PCA, HCA, SVM	[80]
Sepsis	Feces	Fecal headspace	PCA, ROC curve	[81]
Tuberculosis	Breath	Mouthpiece connected to the e-nose	ANN	[82]
Tuberculosis	Breath	Tedlar bag	PCA, k-nn	[83]
Tuberculosis	Breath	Bag	ANN	[84]
Ventilator-associated pneumonia	Breath	Tedlar bag	SVMs, k-nn, ANN, RF, NB	[85]
Ventilator-associated pneumonia	Breath	Ventilator	SVMs, ENN	[86]
Ventilator-associated pneumonia	Breath	Tedlar bag	RF, ROC curve, PCA	[87]
Wound infection	Wound swabs	Bacteria headspace	PLSDA, RBF	[88]
Wound infections	Wound swabs	Bacteria headspace	LDA	[89]

analysis, the technique still does not appear to be ready for widespread clinical use [22]. This is in part due to technical limitations of detection systems comprised of arrays of chemical sensors – in particular difficulties with reproducibility and sensor drift – and in part due to the fact that human breath is a notoriously difficult matrix. It contains thousands of volatile compounds but at very low concentration levels. The exhaled air does not significantly differ in composition from ambient air, and only a fraction of each breath takes part in the gas exchange in the body. While this small fraction does contain volatiles which might be used as indicators of disease, they are present at relatively low concentration levels. For instance, the most abundant volatile organic compound (VOC) present in the expiration of healthy subjects is acetone, and its concentration ranges from approximately 1 to 2 ppm [41]. This likely demarcates the highest practical limit of detection of gas sensors in the e-nose array, and should not be taken for granted when considering the use of low-cost sensors. Furthermore, many sensors, including the prolific metal oxide semiconductor (MOS) sensors, are sensitive to changes in relative humidity – a nontrivial issue when considering direct breath sampling. While this can be to some extent mitigated by using dedicated sampling mechanisms [10] or collecting samples into Tedlar bags, as was done in the majority of reviewed studies (see Fig. 1), it remains an obstacle in developing practical and low-cost solutions for breath analysis. Electronic noses are very sensitive to changes in environmental and sampling conditions, and the real field conditions often differ significantly from the laboratory conditions in which the devices were calibrated. In the case of breath analysis, this might entail the aforementioned issues with relative humidity but also changes in temperature and sampling flow, especially if it relies solely on the subject breathing into the device.

It is also necessary to overcome the issue of sensor drift. While the initial results obtained in numerous test studies are very promising, they do deteriorate rapidly as the sensors' response to a given sample shifts in time due to thermo-mechanical fatigue, poisoning, as well as

adsorption and desorption on the elements of the pneumatic setup [18]. Such a shift in sensor responses invalidates over time the initially trained ML models. Since frequent recalibration of the devices would not be cost-effective, researchers are looking for applications of AI in correcting the sensor drift, as outlined in the following section.

Another issue is that human volatilome is hardly a stable and immutable background for the changing concentrations of possible disease indicators. Indeed, its composition depends on a myriad of long- and short-term factors such as age, sex, diet, habits, time of day, or whether or not the subject had a mild exercise such as ascending a staircase prior to sampling [10, 22]. This further compounds the difficulties with extracting diagnostically useful information from the non-selective responses of the array of chemical sensors and makes the need for the use of elaborate data mining techniques more acute.

Machine Learning in Enhancing the Diagnostic Capability of Electronic Noses

The determination of trace volatile markers of disease is a major challenge. These compounds, mostly volatile organic compounds (VOCs), are present in the expiration at, at best, ppm levels and more commonly at ppb and ppt concentration levels [90]. Analysis of trace compounds is not easy to realize using state-of-the-art instrumental techniques based on mass spectrometry, and so the chemical sensors used for the same task have to be characterized by at least comparable sensitivity to the selected volatile indicators. More than 99% v/v of a breath sample does not carry analytically useful information, and the approximately 6% v/v of H₂O and 5% v/v of CO₂ can produce signal suppression in chemical sensors and matrix effects. This is why it is necessary to process and filter the raw sensor response signals, often using ML techniques, in order to extract information about the subject's condition, as outlined in the previous section. However, interfering substances are only a part of the overall challenge. When

contemplating the implementation of electronic olfaction in medical diagnostics, one has to consider issues specific to chemical sensors, i.e., the noise generated by the sensors during measurement and the long-term signal drift.

Denoising Algorithms

All changes in the sensors' response which are not a direct effect of exposure to the sample and which are caused by the electronic nose lead to noise generation. This in turn negatively impacts the device's detection capabilities, particularly in instances in which the differences between several complex biological samples are minute. Noise can be to some extent reduced using dedicated hardware solutions; however, part of it remains in the form of irreducible noise [91]. This is where signal processing algorithms, including ones based on ML, can be particularly useful. These include median filtering (MF), singular value decomposition (SVD), and PCA for noise reduction [91] or discrete wavelet transform (DWT) [92] among others reported elsewhere [93]. The most recent example of denoising algorithms is the discrete wavelet transform and long short-term memory (DWTLSTM) [94].

A good example of the implementation of these solutions is the use of ML to remove random

noise in the case of electronic noses equipped with an array of optical sensors, in which the removal of spectra noise is crucial for correct sample classification. The most common denoising methods such as Fourier analysis [95], wavelet transform [96], and Savitzky-Golay filtering (SG) [97] often produce very good results; they are, however, not flexible. One of the proposed alternatives is the use of least-squares support vector machine (LSSVM) filtering as a denoising tool for improving optical e-nose performance [98]. The results of its implementation were juxtaposed with other methods, i.e., moving window average (MVA), SG, and wavelet threshold filtering (Fig. 2). Apart from eliminating the random noise, the LSSVM algorithm was capable to preserve the relative extremes of the original spectrum, unlike other methods.

Sensor Drift Compensation

The issue of sensor drift was outlined in the previous section. It represents a major obstacle in the large-scale introduction of standardized, low-cost electronic nose devices for point-of-care diagnostics. Attempts are being made to mitigate its effect through hardware modifications, e.g., through replacing individual sensors [99] or by developing

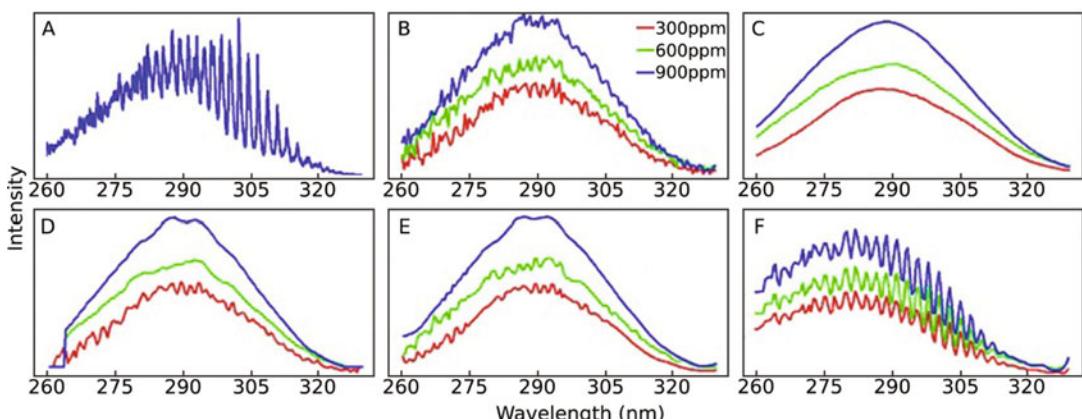


Fig. 2 Performance of denoising algorithms applied to the response signal of an SO_2 optical sensor evaluated based on the normalized correlation coefficient (NCC). The reference spectrum (**a**) was obtained from the HITRAN database. The real spectrum (**b**) was transformed using WMA

(**c**), SG (**d**), wavelet transform (**e**), and LSSVM (**f**) algorithms. Using an ML protocol, it was possible to obtain the highest NCC value of approximately 0.98 regardless of the gas concentration. (Modified from [98], Copyright 2018, with permission from Elsevier)

novel types of carbon nanotube-based sensors which self-compensate for changes in relative humidity [100]. However, compensating for drift at the data processing stage seems to be more practical. The most straightforward approach here would be to recalibrate the sensors using a reference gaseous mixture. This is, however, relatively costly and time-consuming. It is often based on linear response models, which are only applicable in situations in which the drift is linear in nature (e.g., baseline shift) [18]. Another solution is to employ univariate analysis (to the response signal of each of the sensors in the sensor array). Examples include baseline manipulation, frequency analysis, and differential analysis. These methods, while computationally undemanding, require manual calibration. It is also possible to perform multivariate statistics [18], in which the drift compensation is applied not to the response of a single sensor but to the response of the entire sensor matrix [93]. However, all the solutions outlined above require user's involvement – a step which ideally could be bypassed using artificial intelligence.

One solution is to use supervised ML methods such as component correction (CC) or sequential

minimal optimization (SMO). Unfortunately, this entails preparing a training set prior to drift compensation [101]. An alternative is to use semi-supervised methods such as Weighted Geodesic Flow Kernel [102] or multifeature kernel semi-supervised joint learning model (MFKS) [102]. However, the current trend is to try to develop a fully unsupervised approach. Such new developments can now be validated against a common benchmark which allows to juxtapose different models and assess their performance and facilitates the development of novel drift compensation algorithms. This was made possible by the publication of an open dataset of sensor responses prepared by Vergara et al. [103]. It contains 13,910 measurements of 6 gases over a period of 36 months. Another dataset that is commonly used to validate new drift compensation protocols is the CQU sensor drift dataset which contains 1604 measurements using 3 devices, of which 1 was the master device and the 2 remaining ones operating as slave devices, with 5 years of difference between devices of both categories [104]. Selected ML models for sensor drift compensation developed in recent years and validated using the UCI dataset [103] are listed in Table 2

Table 2 The juxtaposition of selected sensor drift compensation methods which were validated using the UCI dataset [103]. Average classification accuracy denotes the percentage of samples from batches 2 to 10 of the dataset,

Type of drift removal method	Method name	Average classification accuracy [%]	Ref.
Component correction	CC-PCA	45.7	[104, 106]
	CC-LDA	49.9	[104, 106, 109]
Label-free correction	MFKS	77.6	[109]
	DRCA	77.6	[104]
	SVM-rbf	38.9	[104]
	SVM-comgfk	64.0	[104, 106, 109]
Extreme learning machine	ELM-rbf	57.9	[106]
	DAELM-T	91.6	[101]
	ODAELM-T	89.1	[101]
Active learning	AL-ACR	84.3	[109]
	AL-DC-ER	93.1	[109]
	AL-ISSMK	89.2	[107]

CC-PCA, component correction PCA; CC-LDA, component correction LDA; MFKS, multifeature kernel semi-supervised joint learning model; DRCA, domain regularized component analysis; SVM-rbf, SVM radial basis function; SVM-comgfk, SVM combined with kernel of geodesic flow kernel; ELM-rbf, ELM radial basis function; DAELM-T, target domain adaptation ELM; ODAELM-T, online target domain adaptation ELM; AL-ACR, AL adaptive confidence rule; AL-DC-ER, AL dynamic clustering error reduction sampling; AL-ISSMK, AL instance-selection strategy on a mixed kernel

containing signals which drifted over time, was correctly classified as the initial signals from batch 1 after the application of the method

together with the corresponding average classification accuracies.

One of the fully unsupervised algorithms for drift correction is the anti-drift which is based on domain regularized component analysis (DRCA) [104]. It projects the data onto a new subspace in such a way that the clusters (more-or-less discrete groups) of the source domain and the target domain have a similar distribution. Another fully unsupervised method is the online target domain (alternatively source domain) adaptation extreme learning machine [101]. Some algorithms assume that the drift is the same for each feature, which is not always the case, especially when the sensor matrix of the electronic nose is comprised of different types of chemical sensors. Selected features can be targeted, e.g., using the discrete binary particle swarm optimization-cosine similarity (DBPSO-CS) classifier which identifies drift-intensive features [105]. It should be noted that the datasets compiled, appended, and continuously concatenated throughout the use of a set of electronic noses developed for a particular diagnostic task might be vastly different depending on the settings, number of analyzed samples, device construction, and numerous other factors. This means that it is difficult to find a versatile drift compensation solution that will be applicable in such scenarios. This issue could be solved using approaches akin to the recently developed discriminative dimensionality reduction algorithm [106] which, owing to the use of two trade-off parameters, can adjust the drift compensation to a given scenario, offering some degree of flexibility. The last 2 years also saw the introduction of methods that are based on the active learning mechanism, including the instance-selection strategy on a mixed kernel (ISSMK) [107] and the active drift calibration sample selection (AC-CSS) [108].

Some ML-based algorithms are aimed at solving more than a single issue. For instance, the multi-classifier tree model can be used for both sensor drift compensation and fixing problems with delayed sensor response. It is a complex approach that also involves the removal of outliers (noise artifacts) using the boxplot approach. The mitigation of the delayed sensor response issue is

achieved by using the sensor response in the transient state (the initial, rapid change of the signal) instead of the steady state, which makes it possible to curtail the time of a single analysis by not waiting for the signal to stabilize [99].

While the use of a single common dataset for the validation of the newly developed drift compensation algorithms certainly facilitates the assessment of their performance, there are drawbacks of using data obtained in strictly controlled settings and for a handful of gaseous substances which might impact the robustness of the models. This is why a recent trend is to supplement the validation protocol with data from own measurements of more complex mixtures of volatiles [108].

Conclusions

At first glance, electronic olfaction seems to be very well suited for the development of low-maintenance, inexpensive tools for noninvasive screening tests and point-of-care diagnostics, thanks to its ability to extract useful information from very complex gaseous mixtures. This ability is realized using artificial intelligence in the form of elaborate machine learning algorithms. However, several methodological issues conspire to prevent the technique from reaching maturity sufficient for widespread adoption in clinical settings. These are related to the complexity of the biological matrices in general, and of breath in particular, and to the hardware limitations of the conventional electronic noses equipped with chemical sensors, such as sensor drift, high limits of detection, and susceptibility to humidity. While some of the difficulties with the realization of electronic nose-based medical diagnostics can be overcome using hardware solutions such as elaborate sampling setups, others, such as sensor drift, can be at least in part alleviated using the recent advances in AI. This is reflected by the recent proliferation of ML algorithms for processing the raw signals and data from sensor arrays and for mining analytically useful information. Even if in the future researchers will abandon the hope of utilizing electronic noses equipped with

low-cost gas sensors in medical diagnostics in favor of their counterparts based on ion mobility spectrometry [110] or mass spectrometry [111], even more will be asked of AI in order to handle the large multivariate datasets and to extract from them diagnostic data while avoiding the curse of dimensionality.

References

- Persaud K, Dodd G. Analysis of discrimination mechanisms in the mammalian olfactory system using a model nose. *Nature* [Internet]. 1982 Sep 23 [cited 2021 Feb 28];299(5881):352–5. <https://doi.org/10.1038/299352a0>
- McEntegart CM, Penrose WR, Strathmann S, Stetter JR. Detection and discrimination of coliform bacteria with gas sensor arrays. *Sensors Actuators B Chem* [Internet]. 2000 Nov [cited 2021 Feb 28];70(1–3):170–6. <http://www.sciencedirect.com/science/article/pii/S092540050000561X>
- Santonico M, Pennazza G, Grasso S, D'Amico A, Bizzarri M. Design and test of a biosensor-based multisensorial system: a proof of concept study. *Sensors (Basel)* [Internet]. 2013 Jan 4 [cited 2021 Feb 28];13(12):16625–40. <http://www.mdpi.com/1424-8220/13/12/16625/htm>
- Shimizu FM, Braunger ML, Riul A, Oliveira ON. Electronic tongues. In: Smart sensors for environmental and medical applications. Wiley; 2020. p. 61–80.
- Shepherd GM. Smell images and the flavour system in the human brain [Internet]. Vol. 444, *Nature*. Nature Publishing Group; [Internet] 2006 [cited 2021 Feb 28]. p. 316–21. <https://www.nature.com/articles/nature05405>
- Röck F, Barsan N, Weimar U. Electronic nose: current status and future trends. *Chem Rev* [Internet]. 2008 [cited 2021 Feb 28];108(2):705–25. http://pubs3.acs.org/acs/journals/doilookup?in_doi=10.1021/cr068121q
- Buck L, Axel R. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell*. 1991;65(1):175–87.
- Marcinkowska R, Namieśnik J, Tobiszewski M. Green and equitable analytical chemistry. Vol. 19, Current opinion in green and sustainable chemistry. Elsevier B.V.; 2019. p. 19–23.
- Majchrzak T, Wojnowski W, Dymerski T, Gębicki J, Namieśnik J. Electronic noses in classification and quality control of edible oils: a review. *Food Chem*. 2018;246:192–201.
- Majchrzak T, Wojnowski W, Piotrowicz G, Gębicki J, Namieśnik J. Sample preparation and recent trends in volatolomics for diagnosing gastrointestinal diseases. *TrAC – Trends Anal Chem*. 2018;108:38–49.
- Wojnowski W, Kalinowska K, Majchrzak T, Płotka-Wasylka J, Namieśnik J. Prediction of the biogenic amines index of poultry meat using an electronic nose. *Sensors*. 2019;19(7):1580.
- Hotel O, Poli JP, Mer-Calfati C, Scorsone E, Saada S. A review of algorithms for SAW sensors e-nose based volatile compound identification. *Sensors Actuators, B: Chem* Elsevier B.V.; [Internet] Feb 1, 2018 cited [2021 Feb 28] p. 2472–82. <https://linkinghub.elsevier.com/retrieve/pii/S0925400517317057>
- Wojnowski W, Dymerski T, Gębicki J, Namieśnik J. Electronic noses in medical diagnostics. *Curr Med Chem*. 2019;26(1):197–215.
- Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice [Internet]. Vol. 338, *BMJ British Medical Journal Publishing Group*; 2009 [cited 2021 Feb 24]. p. 1487–90. <https://www.bmjjournals.org/content/338/bmj.b606>
- Leopold JH, Bos LDJ, Sterk PJ, Schultz MJ, Fens N, Horvath I, et al. Comparison of classification methods in breath analysis by electronic nose. *J Breath Res*. 2015;9(4):046002.
- Marco S, Gutierrez-Galvez A. Signal and data processing for machine olfaction and chemical sensing: a review. *IEEE Sensors J*. 2012;12(11):3189–214.
- Marco S. The need for external validation in machine olfaction: emphasis on health-related applications [Internet]. Vol. 406, *Analytical and bioanalytical chemistry*. Springer; 2014 [cited 2021 Feb 28]. p. 3941–56. <https://link.springer.com/article/10.1007/s00216-014-7807-7>
- Marco S, Gutierrez-Galvez A. Signal and data processing for machine olfaction and chemical sensing: A review. *IEEE Sensors J* [Internet]. 2012 Nov [cited 2021 Feb 28];12(11):3189–214. <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6183455>
- Scott SM, James D, Ali Z. Data analysis for electronic nose systems. *Microchim Acta*. 2006;156(3–4):183–207.
- Wilson AD, Baietto M. Advances in electronic-nose technologies developed for biomedical applications. *Sensors* [Internet]. 2011 Jan 19 [cited 2021 Feb 28];11(12):1105–76. <http://www.mdpi.com/1424-8220/11/1/1105/>
- Bernabei M, Pennazza G, Santonico M, Corsi C, Roscioni C, Paolesse R, et al. A preliminary study on the possibility to diagnose urinary tract cancers by an electronic nose. *Sensors Actuators B Chem*. 2008;131(1):1–4.
- Bikov A, Lázár Z, Horvath I. Established methodological issues in electronic nose research: How far are we from using these instruments in clinical settings of breath analysis? *J Breath Res* [Internet]. 2015 Jun 9 [cited 2021 Feb 28];9(3):034001. <https://doi.org/10.1088/1752-7155/9/3/034001>

23. Palmer CK, Thomas MC, Von Wagner C, Raine R. Reasons for non-uptake and subsequent participation in the NHS Bowel cancer screening programme: a qualitative study. *Br J Cancer* [Internet]. 2014 [cited 2021 Feb 28];110(7):1705–11. <https://www.nature.com/articles/bjc2014125.pdf>
24. Tenero L, Sandri M, Piazza M, Paiola G, Zaffanello M, Piacentini G. Electronic nose: a pilot study to discriminate of children with uncontrolled asthma. *J Breath Res.* 2020;14:046003.
25. Kononov A, Korotetsky B, Jahatsianian I, Gubal A, Vasiliev A, Arsenjev A, et al. Online breath analysis using metal oxide semiconductor sensors (electronic nose) for diagnosis of lung cancer. *J Breath Res.* 2020;14(1):016004.
26. van Velzen P, Brinkman P, Knobel HH, van den Berg JWK, Jonkers RE, Loijmans RJ, et al. Exhaled breath profiles before, during and after exacerbation of COPD: a prospective follow-up study. *COPD J Chronic Obstr Pulm Dis.* 2019;16:330.
27. Schnabel RM, Boumans MLL, Smolinska A, Stobberingh EE, Kaufmann R, Roekaerts PMHJ, et al. Electronic nose analysis of exhaled breath to diagnose ventilator-associated pneumonia. *Respir Med.* 2015;109(11):1454–9.
28. Shafiek H, Fiorentino F, Merino JL, López C, Oliver A, Segura J, et al. Using the electronic nose to identify airway infection during COPD exacerbations. Kostikas K, editor. *PLoS One* [Internet]. 2015 Sep 9 [cited 2021 Feb 11];10(9):e0135199. <https://dx.plos.org/10.1371/journal.pone.0135199>
29. Brinkman P, Wagener AH, Hekking PP, Bansal AT, Maitland-van der Zee AH, Wang Y, et al. Identification and prospective stability of electronic nose (eNose)-derived inflammatory phenotypes in patients with severe asthma. *J Allergy Clin Immunol.* 2019;143(5):1811–1820.e7.
30. Dragonieri S, Quaranta VN, Carratu P, Ranieri T, Resta O. Exhaled breath profiling by electronic nose enabled discrimination of allergic rhinitis and extrinsic asthma. *Biomarkers.* 2019;24(1):70–5.
31. De Vries R, Dagelet YWF, Spoor P, Snoey E, Jak PMC, Brinkman P, et al. Clinical and inflammatory phenotyping by breathomics in chronic airway diseases irrespective of the diagnostic label. *Eur Respir J.* 2018;51(1):1–10.
32. Breathomix. SpiroNose [Internet]. [cited 2021 Feb 24]. <https://www.breathomix.com/spironose-2/>
33. Kim H-Y. Analysis of variance (ANOVA) comparing means of more than two groups. *Restor Dent Endod.* 2014;39(1):74.
34. Bernabei M, Pennazza G, Santonicò M, Corsi C, Roscioni C, Paolesse R, et al. A preliminary study on the possibility to diagnose urinary tract cancers by an electronic nose. *Sensors Actuators, B Chem* [Internet]. 2008 cited [2021 Feb 28];131(1):1–4. <https://doi.org/10.1007/s12149-018-1247-y>
35. Baldini C, Billeci L, Sansone F, Conte R, Domenici C, Tonacci A. Electronic nose as a novel method for diagnosing cancer: a systematic review. *Biosensors* [Internet]. 2020 Jul 25 [cited 2021 Feb 20];10(8):1–21. <https://www.mdpi.com/2079-6374/10/8/84>
36. Wilson AD. Electronic-nose applications in forensic science and for analysis of volatile biomarkers in the human breath. *J Forensic Sci Criminol* [Internet]. 2014 [cited 2021 Feb 28];1(1):1–21. https://www.srs.fs.usda.gov/pubs/ja/2014/ja_2014_wilson_001.pdf
37. Wilson AD. Application of electronic-nose technologies and VOC-biomarkers for the noninvasive early diagnosis of gastrointestinal diseases. *Sensors (Switzerland).* 2018;18(8):2613.
38. Wilson AD. Recent applications of electronic-nose technologies for the noninvasive early diagnosis of gastrointestinal diseases†. *Proceedings.* 2017;2 (3):147.
39. Wilson AD. Applications of electronic-nose technologies for noninvasive early detection of plant, animal and human diseases. *Chemosensors.* 2018;6(4):1–36.
40. Fitzgerald JE, Bui ETH, Simon NM, Fenniri H. Artificial nose technology: status and prospects in diagnostics. *Trends Biotechnol* [Internet]. 2017 Jan [cited 2021 Feb 28];35(1):33–42. <https://doi.org/10.1016/j.tibtech.2016.08.005>
41. Anderson JC. Measuring breath acetone for monitoring fat loss: review [Internet]. Vol. 23, *Obesity*. Blackwell Publishing, 2015 [cited 2021 Feb 25]. p. 2327–34. <https://doi.org/10.1002/oby.21242>.
42. Guo D, Zhang D, Li N, Zhang L, Yang J. A novel breath analysis system based on electronic olfaction. *IEEE Trans Biomed Eng.* 2010;57(11):2753–63.
43. Dragonieri S, Quaranta VN, Carratu P, Ranieri T, Resta O. Exhaled breath profiling by electronic nose enabled discrimination of allergic rhinitis and extrinsic asthma. *Biomarkers* [Internet]. 2019 [cited 2021 Feb 28];24(1):70–5. <https://doi.org/10.1080/1354750X.2018.1508307>
44. Dragonieri S, Quaranta VN, Carratu P, Ranieri T, Marra L, D'Alba G, et al. An electronic nose may sniff out amyotrophic lateral sclerosis. *Respir Physiol Neurobiol* [Internet]. 2016 [cited 2021 Feb 28];232: 22–5. <https://doi.org/10.1016/j.resp.2016.06.005>
45. De Heer K, Kok MGM, Fens N, Weersink EJM, Zwinderen AH, Van Der Schee MPC, et al. Detection of airway colonization by *Aspergillus fumigatus* by use of electronic nose technology in patients with cystic fibrosis (*Journal of Clinical Microbiology* (2016) 54:3 (569–575)). *J Clin Microbiol.* 2016;54 (7):1926.
46. Ibrahim B, Basanta M, Cadden P, Singh D, Douce D, Woodcock A, et al. Non-invasive phenotyping using exhaled volatile organic compounds in asthma. *Thorax.* 2011;66(9):804–9.
47. Bannier MAGE, Van De Kant KDG, Jöbsis Q, Dompeeling E. Feasibility and diagnostic accuracy of an electronic nose in children with asthma and cystic fibrosis. *J Breath Res.* 2019;13(3):036009.

48. Brinkman P, van de Pol M, Gerritsen M, Bos L, Dekker T, Smids B, et al. Exhaled breath profiles in the monitoring of loss of control and clinical recovery in asthma. *Clin Exp Allergy*. 2017;47:1159.
49. De Vries R, Dagelet YWF, Spoor P, Snoey E, Jak PMC, Brinkman P, et al. Clinical and inflammatory phenotyping by breathomics in chronic airway diseases irrespective of the diagnostic label. *Eur Respir J* [Internet]. 2018 [cited 2021 Feb 28];51(1):1–10. <https://doi.org/10.1183/13993003.01817-2017>
50. Saidi T, Zaim O, Moufid M, El Bari N, Ionescu R, Bouchikhi B. Exhaled breath analysis using electronic nose and gas chromatography-mass spectrometry for non-invasive diagnosis of chronic kidney disease, diabetes mellitus and healthy subjects. *Sensors Actuators, B Chem* [Internet]. 2018 [cited 2021 Feb 28];257:178–88. <https://doi.org/10.1016/j.snb.2017.10.178>
51. Westenbrink E, Arasaradnam RP, O'Connell N, Bailey C, Nwokolo C, Bardhan KD, et al. Development and application of a new electronic nose instrument for the detection of colorectal cancer. *Biosens Bioelectron* [Internet]. 2015 [cited 2021 Feb 28];67:733–8. <https://doi.org/10.1016/j.bios.2014.10.044>
52. Finamore P, Pedone C, Scarlata S, Di Paolo A, Grasso S, Santonicò M, et al. Validation of exhaled volatile organic compounds analysis using electronic nose as index of COPD severity. *Int J COPD*. 2018;13:1441–8.
53. Wintjens AGWE, Hintzen KFH, Engelen SME, Lubbers T, Savelkoul PHM, Wesseling G, et al. Applying the electronic nose for pre-operative SARS-CoV-2 screening. *Surg Endosc* [Internet]. 2020 [cited 2021 Feb 28];(0123456789). <https://doi.org/10.1007/s00464-020-08169-0>
54. Shan B, Broza YY, Li W, Wang Y, Wu S, Liu Z, et al. Multiplexed nanomaterial-based sensor Array for detection of COVID-19 in exhaled breath. *ACS Nano*. 2020;14(9):12125–32.
55. Schuermans VNE, Li Z, Jongen ACHM, Wu Z, Shi J, Ji J, et al. Pilot study: detection of gastric cancer from exhaled air analyzed with an electronic nose in Chinese patients. *Surg Innov*. 2018;25(5):429–34.
56. van de Goor RMGE, Leunis N, van Hooren MRA, Francisca E, Masclee A, Kremer B, et al. Feasibility of electronic nose technology for discriminating between head and neck, bladder, and colon carcinomas. *Eur Arch Oto-Rhino-Laryngol*. 2017;274(2):1053–60.
57. van Hooren MRA, Leunis N, Brandsma DS, Dingemans AMC, Kremer B, Kross KW. Differentiating head and neck carcinoma from lung carcinoma with an electronic nose: a proof of concept study. *Eur Arch Oto-Rhino-Laryngol*. 2016;273(11):3897–903.
58. Van Geffen WH, Bruins M, Kerstjens HAM. Diagnosing viral and bacterial respiratory infections in acute COPD exacerbations by an electronic nose: a pilot study. *J Breath Res*. 2016;10(3):036001.
59. Plaza V, Crespo A, Giner J, Merino JL, Ramos-Barbón D, Mateus EF, et al. Inflammatory asthma phenotype discrimination using an electronic nose breath analyzer. *J Investig Allergol Clin Immunol* [Internet]. 2015 [cited 2021 Feb 28];25(6):431–7. <http://europemc.org/abstract/MED/26817140>
60. Tiele A, Wicaksno A, Kansara J, Arasaradnam RP, Covington JA. Breath analysis using enose and ion mobility technology to diagnose inflammatory bowel disease – a pilot study. *Biosensors*. 2019;9(2):1–15.
61. Moor CC, Oppenheimer JC, Nakshbandi G, Aerts JGJV, Brinkman P, Maitland-Van Der Zee AH, et al. Exhaled breath analysis by use of eNose technology: a novel diagnostic tool for interstitial lung disease. *Eur Respir J*. 2021;57(1):2002042.
62. van de Goor RMGE, Hardy JCA, van Hooren MRA, Kremer B, Kross KW. Detecting recurrent head and neck cancer using electronic nose technology: a feasibility study. *Head Neck*. 2019;41(9):2983–90.
63. Huang CH, Zeng C, Wang YC, Peng HY, Lin CS, Chang CJ, et al. A study of diagnostic accuracy using a chemical sensor array and a machine learning technique to detect lung cancer. *Sensors (Switzerland)*. 2018;18(9):2845.
64. Lu B, Fu L, Nie B, Peng Z, Liu H. A novel framework with high diagnostic sensitivity for lung cancer detection by electronic nose. *Sensors (Switzerland)*. 2019;19(23):1–29.
65. Gasparri R, Santonicò M, Valentini C, Sedda G, Borri A, Petrella F, et al. Volatile signature for the early diagnosis of lung cancer. *J Breath Res* [Internet]. 2016 Feb 9 [cited 2021 Feb 11];10(1):016007. <https://iopscience.iop.org/article/10.1088/1752-7155/10/1/016007>
66. Tirzīte M, Bukovskis M, Strazda G, Jurka N, Taivans I. Detection of lung cancer in exhaled breath with an electronic nose using support vector machine analysis. *J Breath Res*. 2017;11(3):036009.
67. van de Goor R, van Hooren M, Dingemans AM, Kremer B, Kross K. Training and validating a portable electronic nose for lung cancer screening. *J Thorac Oncol* [Internet]. 2018 [cited 2021 Feb 28];13(5):676–81. <https://doi.org/10.1016/j.jtho.2018.01.024>
68. McWilliams A, Beigi P, Srinidhi A, Lam S, MacAulay CE. Sex and smoking status effects on the early detection of early lung cancer in high-risk smokers using an electronic nose. *IEEE Trans Biomed Eng*. 2015;62(8):2044–54.
69. Lamote K, Brinkman P, Vandermeersch L, Vynck M, Sterk PJ, Van Langenhove H, et al. Breath analysis by gas chromatography-mass spectrometry and electronic nose to screen for pleural mesothelioma: a cross-sectional case-control study. *Oncotarget*. 2017;8(53):91593–602.
70. De Meij TGJ, Van Der Schee MPC, Berkhout DJC, Van De Velde ME, Jansen AE, Kramer BW, et al. Early detection of necrotizing enterocolitis by fecal volatile organic compounds analysis. *J Pediatr*

- [Internet]. 2015 [cited 2021 Feb 28];167(3):562–567. e1. <https://doi.org/10.1016/j.jpeds.2015.05.044>
71. Dragonieri S, Porcelli F, Longobardi F, Carratu P, Aliani M, Ventura VA, et al. An electronic nose in the discrimination of obese patients with and without obstructive sleep apnoea. *J Breath Res* [Internet]. 2015 Jun 1 [cited 2021 Feb 11];9(2):026005. <https://iopscience.iop.org/article/10.1088/1752-7155/9/2/026005>
72. Peters Y, Schrauwen RWM, Tan AC, Bogers SK, De Jong B, Siersma PD. Detection of Barrett's oesophagus through exhaled breath using an electronic nose device. *Gut*. 2020;69(7):1169–72.
73. Dragonieri S, Quaranta VN, Carratu P, Ranieri T, Resta O. Exhaled breath profiling in patients with COPD and OSA overlap syndrome: a pilot study. *J Breath Res* [Internet]. 2016 Nov 3 [cited 2021 Feb 11];10(4):041001. <https://iopscience.iop.org/article/10.1088/1752-7155/10/4/041001>
74. Finberg JPM, Schwartz M, Jeries R, Badarny S, Nakhleh MK, Abu Daoud E, et al. Sensor array for detection of early stage Parkinson's disease before medication. *ACS Chem Neurosci* [Internet]. 2018 [cited 2021 Feb 28];9(11). <https://doi.org/10.1021/acscnemo.8b00245>
75. Cavaleiro Rufo J, Paciência I, Mendes FC, Farraia M, Rodolfo A, Silva D, et al. Exhaled breath condensate volatileome allows sensitive diagnosis of persistent asthma. *Allergy Eur J Allergy Clin Immunol*. 2019;74(3):527–34.
76. Yang H-Y, Peng H-Y, Chang C-J, Chen P-C. Diagnostic accuracy of breath tests for pneumoconiosis using an electronic nose. *J Breath Res* [Internet]. 2017 Nov 29 [cited 2021 Feb 28];12(1):016001. <https://iopscience.iop.org/article/10.1088/1752-7163/aa857d>
77. Suarez-Cuartin G, Giner J, Merino JL, Rodriguez-Troyano A, Feliu A, Perea L, et al. Identification of *Pseudomonas aeruginosa* and airway bacterial colonization by an electronic nose in bronchiectasis. *Respir Med* [Internet]. 2018 [cited 2021 Feb 28];136 (December 2017):111–7. <https://doi.org/10.1016/j.rmed.2018.02.008>
78. Brekelmans M, Fens N, Brinkman P, Bos L, Sterk P, Gerlag D. Smelling the diagnosis: the electronic nose as diagnostic tool in inflammatory arthritis. A case-reference study. *PLoS One* [Internet]. 2016 [cited 2021 Feb 28];11. <https://doi.org/10.1371/journal.pone.0151715>
79. De Vries R, Muller M, Van Der Noort V, Theelen WSME, Schouten RD, Hummelink K, et al. Prediction of response to anti-PD-1 therapy in patients with non-small-cell lung cancer by electronic nose analysis of exhaled breath. *Ann Oncol* [Internet]. 2019 [cited 2021 Feb 28];30(10):1660–6. <https://doi.org/10.1093/annonc/mdz279>
80. Saidi T, Tahri K, El Bari N, Ionescu R, Bouchikhi B. Detection of seasonal allergic rhinitis from exhaled breath VOCs using an electronic nose based on an array of chemical sensors. *2015 IEEE Sensors – Proc*. 2015. p. 1–4.
81. Berkhouit DJC, Niemarkt HJ, Buijck M, Van Weissenbruch MM, Brinkman P, Benninga MA, et al. Detection of sepsis in preterm infants by fecal volatile organic compounds analysis: a proof of principle study. *J Pediatr Gastroenterol Nutr*. 2017;65(3):e47–52.
82. Coronel Teixeira R, Rodríguez M, Jiménez de Romero N, Bruins M, Gómez R, Yntema JB, et al. The potential of a portable, point-of-care electronic nose to diagnose tuberculosis. *J Infect* [Internet]. 2017 [cited 2021 Feb 28];75(5):441–7. <https://doi.org/10.1016/j.jinf.2017.08.003>
83. Zetola NM, Modongo C, Matsiri O, Tamuhla T, Mbongwe B, Matlhagela K, et al. Diagnosis of pulmonary tuberculosis and assessment of treatment response through analyses of volatile compound patterns in exhaled breath samples. *J Infect* [Internet]. 2017 [cited 2021 Feb 28];74(4):367–76. <https://doi.org/10.1016/j.jinf.2016.12.006>
84. Mohamed EI, Mohamed MA, Moustafa MH, Abdel-Mageed SM, Moro AM, Baess AI, et al. Qualitative analysis of biological tuberculosis samples by an electronic nose-based artificial neural network. *Int J Tuberc Lung Dis*. 2017;21(7):810–7.
85. Chen CY, Lin WC, Yang HY. Diagnosis of ventilator-associated pneumonia using electronic nose sensor array signals: solutions to improve the application of machine learning in respiratory research. *Respir Res*. 2020;21(1):1–12.
86. Liao YH, Wang ZC, Zhang FG, Abbad MF, Shih CH, Shieh JS. Machine learning methods applied to predict ventilator-associated pneumonia with *Pseudomonas aeruginosa* infection via sensor array of electronic nose in intensive care unit. *Sensors (Switzerland)*. 2019;19(8):1866.
87. Schnabel RM, Boumans MLL, Smolinska A, Stobberingh EE, Kaufmann R, Roekaerts PMHJ, et al. Electronic nose analysis of exhaled breath to diagnose ventilator-associated pneumonia. *Respir Med* [Internet]. 2015 [cited 2021 Feb 28];109(11):1454–9. <https://doi.org/10.1016/j.rmed.2015.09.014>
88. He P, Pengfei J, Qiao S, Duan S. Self-taught learning based on sparse autoencoder for E-nose in wound infection detection. *Sensors (Switzerland)*. 2017;17 (10):2279.
89. Saviauk T, Kiiski JP, Nieminen MK, Tamminen NN, Roine AN, Kumpulainen PS, et al. Electronic nose in the detection of wound infection bacteria from bacterial cultures: a proof-of-principle study. *Eur Surg Res*. 2018;59:1–11.
90. Smith D, Španěl P. The challenge of breath analysis for clinical diagnosis and therapeutic monitoring. *Analyst* [Internet]. 2007 Apr 30 [cited 2021 Feb 28];132(5):390–6. <http://xlink.rsc.org/?DOI=B700542N>
91. Jha SK, Yadava RDS. Performance assessment of PCA, MF and SVD methods for denoising in

- chemical sensor array based electronic nose system. Sensors Transducers [Internet]. 2011 [cited 2021 Feb 24];129(6):57–68. <http://www.sensorsportal.com>
92. Wijaya DR, Sarno R, Zulaika E. Noise filtering framework for electronic nose signals: an application for beef quality monitoring. Comput Electron Agric [Internet]. 2019 Feb 1 [cited 2021 Feb 23];157:305–21. www.elsevier.com/locate/compag
93. Distante C, Leo M, Siciliano P, Persaud KC. On the study of feature extraction methods for an electronic nose. Sensors Actuators B Chem. 2002;87(2):274–88.
94. Wijaya DR, Sarno R, Zulaika E. DWTLSTM for electronic nose signal processing in beef quality monitoring. Sensors Actuators B Chem. 2021;326: 128931.
95. Yatabe K, Oikawa Y. Convex optimization-based windowed Fourier filtering with multiple windows for wrapped-phase denoising. Appl Opt [Internet]. 2016 Jun 10 [cited 2021 Feb 27];55(17):4632. <https://doi.org/10.1364/AO.55.004632>
96. Arboleda C, Wang Z, Stampaconi M. Wavelet-based noise-model driven denoising algorithm for differential phase contrast mammography. Opt Express [Internet]. 2013 May 6 [cited 2021 Feb 27];21(9):10572. <https://www.osapublishing.org/viewmedia.cfm?uri=oe-21-9-10572&seq=0&html=true>
97. Agarwal S, Rani A, Singh V, Mittal AP. EEG signal enhancement using cascaded S-Golay filter. Biomed Signal Process Control. 2017;36:194–204.
98. Zhang W, Tian F, Song A, Hu Y. Research on an optical e-nose denoising method based on LSSVM. Optik (Stuttg). 2018;168:118–26.
99. Rehman A ur, Belhaouari SB, Ijaz M, Bermak A, Hamdi M. Multi-classifier tree with transient features for drift compensation in electronic nose. IEEE Sensors J [Internet]. 2020 [cited 2021 Feb 28]; http://www.ieee.org/publications_standards/publications/rights/index.html
100. Falco A, Loghin FC, Becherer M, Lugli P, Salmerón JF, Rivadeneyra A. Low-cost gas sensing: dynamic self-compensation of humidity in CNT-based devices. ACS Sensors [Internet]. 2019 [cited 2021 Feb 23];4 (12):3141–6. <https://pubs.acs.org/sharingguidelines>
101. Ma Z, Luo G, Qin K, Wang N, Niu W. Online sensor drift compensation for E-nose systems using domain adaptation and extreme learning machine. Sensors (Switzerland) [Internet]. 2018 Mar 1 [cited 2021 Feb 24];18(3):742. <http://www.mdpi.com/1424-8220/18/3/742>
102. Liu Q, Li X, Ye M, Ge SS, Du X. Drift compensation for electronic nose by semi-supervised domain adaptation. IEEE Sensors J. 2014;14(3):657–65.
103. Vergara A, Vembu S, Ayhan T, Ryan MA, Homer ML, Huerta R. Chemical gas sensor drift compensation using classifier ensembles. Sensors Actuators, B Chem [Internet]. 2012 [cited 2021 Feb 28];166–167: 320–9. <https://doi.org/10.1016/j.snb.2012.01.074>
104. Zhang L, Liu Y, He Z, Liu J, Deng P, Zhou X. Anti-drift in E-nose: a subspace projection approach with drift reduction. Sensors Actuators B Chem. 2017;253: 407–17.
105. Rehman AU, Bermak A. Drift-insensitive features for learning artificial olfaction in E-nose system. IEEE Sensors J. 2018;18(17):7173–82.
106. Yi Z. Discriminative dimensionality reduction for sensor drift compensation in electronic nose: a robust, low-rank, and sparse representation method. Expert Syst Appl. 2020;148:113238.
107. Liu T, Li D, Chen Y, Wu M, Yang T, Cao J. Online drift compensation by adaptive active learning on mixed kernel for electronic noses. Sensors Actuators B Chem. 2020;316:128065.
108. Liu T, Li D, Chen J. An active method of online drift-calibration-sample formation for an electronic nose. Meas J Int Meas Confed. 2021;171:108748.
109. Liu T, Li D, Chen J, Chen Y, Yang T, Cao J. Active learning on dynamic clustering for drift compensation in an electronic nose system. Sensors (Switzerland) [Internet]. 2019 Aug 19 [cited 2021 Feb 24];19(16):3601. <https://www.mdpi.com/1424-8220/19/16/3601>
110. Steinbach J, Goedicke-Fritz S, Tutdibi E, Stutz R, Kaiser E, Meyer S, et al. Bedside measurement of volatile organic compounds in the atmosphere of neonatal incubators using ion mobility spectrometry. Front Pediatr [Internet]. 2019 Jun 18 [cited 2021 Feb 28];7:4–8. <https://www.frontiersin.org/article/10.3389/fped.2019.00248/full>
111. Casas-Ferreira AM, Nogal-Sánchez M del, Pérez-Pavón JL, Moreno-Cordero B. Non-separative mass spectrometry methods for non-invasive medical diagnostics based on volatile organic compounds: a review. Anal Chim Acta [Internet]. 2019 Jan [cited 2021 Feb 28];1045:10–22. <https://linkinghub.elsevier.com/retrieve/pii/S0003267018308560>



Artificial Intelligence in Telemedicine

87

Jefferson Gomes Fernandes

Contents

Introduction	1220
The Basics of Telemedicine	1220
The Integration of Artificial Intelligence in Telemedicine	1222
Telemedicine and Artificial Intelligence	1223
Teleophthalmology and AI	1223
Telestroke and AI	1224
Teledermatology and AI	1225
Telemedicine, Artificial Intelligence, and Education	1225
References	1226

Abstract

Telemedicine can be defined as the practice of medicine, between different places, through

Supplementary Information: The online version of this chapter (https://doi.org/10.1007/978-3-030-64573-1_93) contains supplementary material, which is available to authorized users.

J. G. Fernandes (✉)

Telemedicine Education Program, Paulista Medical Association, São Paulo, Brazil

Brazilian Association of Telemedicine and Telehealth, Rio de Janeiro, Brazil

Education Program, International Society for Telemedicine and eHealth, Geneva, Switzerland

CEO, SPECIS – a Health Consulting Firm, São Paulo, Brazil
e-mail: jeffernandes@specismed.com

the responsible use of information and communication technologies. From a steady pace of progressive expansion, telemedicine has grown exponentially worldwide due to the emergence of the COVID-19 pandemic. The addition of artificial intelligence to telemedicine can expand and improve its capabilities, giving endless possibilities for developing solutions for specific healthcare needs. AI in telemedicine can make a significant contribution for the implementation of the continuum of healthcare and promote and facilitate greater access to integrated healthcare, where and when necessary. The potential impact of AI in telemedicine can be identified around four emergent trends: patient monitoring, healthcare information technology, intelligent assistance and diagnosis, and information analysis collaboration. However, their implementation

in healthcare faces safety, ethical, efficacy, efficiency, regulatory, and financial challenges. Their adoption will increase if physicians serve as knowledgeable and supportive guides and leaders in the process. The evidence of the benefit of AI will need to be convincing enough for the medical community and the patients to adopt the technology. For physicians AI can be helpful in aiding in decision-making and for specific tasks to improve healthcare delivery. It is also of extreme value by automating administrative tasks to free up valuable time that can be dedicated to the direct care of patients. AI-enabled telemedicine should suit with existing clinical practice. It requires a framework that could be based on technical and clinical considerations, reliability, reproducibility, usability, accessibility, and costs. This chapter also discusses their benefits and limitations in several medical specialties.

Keywords

Telemedicine · Artificial intelligence · Medicine · Physicians · Machine learning · Deep learning · Education · Benefits · Limitations

Introduction

Telemedicine is considered one of the main innovations in healthcare. It increases the accessibility to healthcare services and improves the quality of medical care and organizational efficiency [1].

The addition of artificial intelligence (AI) to telemedicine can expand and improve its capabilities giving endless possibilities for development of solution of specific needs. Both can assist doctors in providing better quality healthcare services to patients. The association of AI with telemedicine can improve health outcomes and enhance patient's experience. They can also improve and speed screening and diagnosis of diseases, make diagnosis more precise and personalized, and reduce patients in person visits.

AI in telemedicine can make a significant contribution to the implementation of the continuum of healthcare. They can promote and facilitate

greater access to integrated healthcare, where and when necessary throughout the citizen's life cycle. The management of chronic diseases, for instance, needs a multidisciplinary continuous and coordinated care. The consistent communications and connections across different elements of healthcare delivery can be supported by remote care. AI can help address this need, by enabling the intelligent information and communication environment in which health professionals could interact and by providing a knowledge base for patient's management [2].

The Basics of Telemedicine

The World Health Organization defined telemedicine as “the delivery of healthcare services, where distance is a critical factor, by all healthcare professionals using information and communication technologies for the exchange of valid information for diagnosis, treatment and prevention of disease and injuries, research and evaluation, and for the continuing education of healthcare providers, all in the interests of advancing the health of individuals and their communities.” [3]

Despite having been published approximately 22 years ago, this definition continues to be used, and some of its components might need to be reviewed such as “where distance is a critical factor.” The use and experience gained by patients, doctors, and health services demonstrate that even patients who are close to health centers can benefit from telemedicine. Each time more the care of people in their homes is being valued, due to the comfort and safety they bring, especially in times of SARS-CoV-2 pandemic.

Although in several publications telemedicine and telehealth are considered synonymous and used interchangeably [4], technical and regulatory aspects and specificity of various health professions indicate that they should be considered as different. For this chapter, we will distinguish telemedicine from telehealth with the former restricted to service delivery by physicians only and the latter signifying services provided by health professionals in general, such as tele-nursing, telenutrition, telepsychology, and others.

We propose a simpler but no less efficient definition of telemedicine: “Telemedicine is the practice of medicine, between different places, through the responsible use of information and communication technologies.”

By responsible use of telemedicine, we mean the need for its practice to be ethical (in the same way as we practice in face-to-face care) and safe, with privacy and confidentiality (for patient’s protection) and with quality (seeking the best possible patient’s outcomes).

Several new terms are being used to bring a broader and holistic view of people’s healthcare in a digital era, such as digital health, eHealth, virtual care, connected care, and connected health. They might have small differences in their meanings and so should be understood in the contexts in which they are used. One can consider that telemedicine is inserted within these denominations.

Digital health can be simplistically defined as the use of digital technologies for health. It encompasses advanced computing sciences in big data and artificial intelligence. Mobile health (mHealth) is defined as “the use of mobile wireless technologies for health.” [5]

Telemedicine can be practiced synchronously or asynchronously (store-and-forward). Among its modalities we can list the following:

- Teletriage – assessment of symptoms for referral of the patient to the care they need.
- Teleconsultation – physician direct to patient or interaction between physicians.
- Telediagnosis – transmission and registration of complementary exams of the patient to be interpreted by specialists.
- Telemonitoring – continuous collection of physiological parameters, transmitted to a provider for diagnostic and therapeutic purposes.

These terms and definitions vary between countries. Remote patient monitoring (telemonitoring) includes transmission, registration, processing of body parameters, and medical management through electronic systems. It can be performed using wireless devices and wearable or implantable sensors. It allows continuous support and management of chronic diseases and can be synchronous or asynchronous, depending on

the patient’s needs. The application of AI and machine learning can allow better disease surveillance and early detection, incorporate and coordinate data from additional tools like location finders (GPS), improve diagnostics, and support for clinical decision.

It should be emphasized that telemedicine is a method of healthcare. It is not a tool. Tools are hardware, software, Internet, and others that allow technical support for the practice of telemedicine. As a method of healthcare, it should be learned by physicians through a qualified training process to allow it to be practiced safely and with quality. Telemedicine also has limitations, and doctors must know about them together with the specific characteristics of its practice in different medical specialties.

There is ample evidence in the scientific and real-world literature confirming the safety, efficacy, and cost-effectiveness of telemedicine [6–8]. The main benefits are the following:

- Contributes to the digital transformation strategy
- Expands access and brings citizens closer to health services
- Increases healthcare efficacy and effectiveness
- Contributes to improve clinical outcomes
- Increases patient involvement and self-care
- Strengthens coordination of care
- Assists in resolving geographic and social health inequalities
- Allows continuous and articulated monitoring between different levels of care
- Contributes to the organization of health systems
- Reduces health costs

The benefits of telemedicine have been demonstrated in several medical specialties. A meta-analysis of the effectiveness of telemedicine in diabetes mellitus showed that telemedicine interventions were more effective than usual care specially in managing type 2 diabetes [9].

In cardiovascular patients remote monitoring is cost-effective and safe, and it decreases the staff workload and improves the prognosis of heart failure patients [10].

In teleoncology, several effective interventions were identified: telepathology, remote chemotherapy supervision, symptom management, survivorship, and palliative care [11].

However, despite these benefits, telemedicine represented a small fraction of all healthcare activities and spending. Barriers to its development include regulatory uncertainty, uneven financing and reimbursement, and unclear governance.

This scenario seems to be changing. As a clear evolution of the use of digital health and telemedicine, the recent Digital Healthcare Act in Germany has made possible for some health applications such as apps and web-based programs to be prescribed by doctors with the costs covered by the statutory health insurance.

Moreover, with the emergence of the COVID-19 pandemic, there was an exponential growth in the use of telemedicine, and it was possible to explore and demonstrate all the potential benefits of this method of healthcare. Through telemedicine, people can receive a wide range of medical services at home, overcoming the distance barrier. It has been possible to remotely screen patients with symptoms, avoid unnecessary visits to health services, and assist symptomatic patients who can be cared for at home. By keeping potentially infected individuals out of health services, there is a reduction in the risk of contamination from other people, doctors, and health teams.

Of great importance has also been the care of people with chronic diseases, when face-to-face visits are not necessary, especially the elderly, who have a high risk of complications and death when contaminated by SARS-CoV-2.

Hospitals and doctors more experienced in handling serious cases at COVID-19 can support colleagues in multiple hospitals, anywhere in the country or even from other countries [12].

The Integration of Artificial Intelligence in Telemedicine

Among the most commonly used definitions for AI two can be highlighted: (1) computer systems that perform tasks normally requiring human intelligence and (2) AI as “augmented

intelligence”: computer algorithms designed to enhance the capabilities of highly trained professionals.

In medicine, AI is expected to address the high rate of avoidable medical errors and workflow and delivery inefficiencies associated with contemporary healthcare provision [13].

The potential impact of AI in telemedicine can be identified around four emergent trends: patient monitoring, healthcare information technology, intelligent assistance and diagnosis, and information analysis collaboration [14].

Despite these impacts the integration of AI in telemedicine into everyday practice has been limited. Their implementation in healthcare faces safety, ethical, efficacy, efficiency, regulatory, and financial challenges. Their adoption will increase if physicians serve as knowledgeable and supportive guides and leaders in the process [15]. The evidence of the benefit of AI will need to be convincing enough for the medical community and the patients to adopt the technology [16].

AI may be proven to be more accurate than a physician and detect features humans cannot in specific situations. However, one cannot expect that algorithms will be able to replace the comprehensive role of physicians.

Physicians are trained to react when more information from a patient becomes available and make new diagnosis and therapeutic decisions once needed, in a dynamic approach. AI gives its opinion based on “engineered” data as input, and whatever happens to the patients afterward is disconnected from the AI’s original opinion increasing the risk of errors/mistakes. Physicians need to understand what is behind the AI proposed diagnosis to understand what the possible consequences or unintended side effects are. AI is unable yet to diagnose and manage patients without physician input. At present, it has an assistive role with clinician oversight and ultimate responsibility.

For physicians AI can be helpful in aiding in decision-making and for specific tasks to improve healthcare delivery. It is also of extreme value by automating administrative tasks, through its smart algorithms, to free up valuable time that can be dedicated to the direct care of patients [17] as

approximately 50% of a physician's time is spent working on electronic health records [18]. An innovation that could help physicians is speech recognition in replacing the use of keyboards to enter and retrieve information.

The role of history taking in the initial diagnosis of a patient's health problem is higher than patient's physical examination. Some studies had reported a contribution of 76% by history and 11% by examination [19]. As of today, the art of history taking and examination are getting compromised by the increased availability and use mainly of imaging tests such as computed tomography and magnetic resonance. History taking consumes much of a physician's time, so it tends not to be used to the extent feasible via telemedicine. AI as a part of a telehealth application could help in the diagnosis process by providing prompts and clues and suggesting the right next questions based on the answer, thus saving the doctor's time [20].

To be scalable AI-enabled telemedicine should suit existing clinical practice. Real-world validation requires a rigorous and transparent framework, flexible enough to be applied to a broad range of technological innovation. Such a framework could be based on technical and clinical considerations, reliability, reproducibility, usability, accessibility, and cost [21]. It should also include an independent governance and ethical aspects.

AI can be effectively advantageous when it comes to analytical reasoning and problem-solving, particularly when large amounts of data are involved. However, AI should not be overutilized and disregard the vital physician skills in healthcare [20]. AI is unable to fully take and interpret contextual information or determine between relevant versus nonrelevant informational input. Physicians are required to deal with the contingencies of healthcare. An AI-powered system should prioritize the collective good and performance based on evidence-based management principles [20].

Patient data are sensitive in nature and subjected to numerous regulations and legislations such as HIPAA and GDPR to ensure that patient data are collected and stored in compliance with security norms. Adequate measures are needed,

which also include data encryption and blockchain, which add a layer of security in relation to storage and sharing of healthcare data.

While regulatory bodies attempt to provide guidance for users and payers for new technologies, they are challenged by the rapid change of pace of AI and its use in telemedicine. The legal aspects of AI adoption need to be better clarified. The American Medical Association (AMA) advocates that in case of system failure or misdiagnosis from autonomous AI systems, the developers must take liability and maintain their own medical liability insurance with their users [22].

The cost of AI embedded in telemedicine is also a potential barrier to adoption. The heterogeneity in the insurance policy and medicolegal regulations are key challenges for its clinical implementation.

Telemedicine and Artificial Intelligence

Teleophthalmology and AI

Individuals diagnosed with diabetes mellitus need regular and repetitive annual retinal screening for detection and treatment of diabetic retinopathy. Diabetic retinopathy is an increasing problem and early screening, and timely treatment can reduce the burden of sight threatening. Such screening is done by fundus examination by ophthalmologists or by color fundus photography using conventional fundus cameras. In teleophthalmology programs, the digital retinal images are sent to a centralized reading center for assessment for the presence of DR.

The lack of specialized professionals in relation to the demand of patients who should undergo this exam could be solved with the provision of an automated imaging system. Using AI and machine learning, a subfield of AI, with devices within easy reach of the patient, the diagnosis can be made by teleophthalmology. AI systems embedded in the screening process could analyze retinal images and compare them with previous samples to determine the stage of diabetic retinopathy.

However, the available automated imaging systems and devices still have sensitivity and specificity below 90%, so false-positive and false-negative results might occur. The dilated eye examination remains the gold standard of screening. Once proven to be sensitive and specific enough, AI systems can change screening programs and population-based teleophthalmology. A comprehensive practical guidance on telemedicine in diabetic retinopathy has been recently published by the American Telemedicine Association Ocular Telehealth Special Interest Group [23].

Deep learning has also been used in teleophthalmology for the screening and diagnosis of other vision diseases, such as glaucoma and age-related macular degeneration [24].

Telestroke and AI

Stroke is one of the main causes of mortality and disability worldwide. Several studies have shown that stroke has a huge global impact with a considerable financial cost for both health and care services and to patients and their families. Stroke mortality varies considerably according to social and economic development, and approximately 85% of strokes occur in low- and middle-income countries with one-third affecting the economically active population. The demographic transition occurring in most developing countries toward an increase in the older population will amplify the impact of stroke [25].

Delays in diagnosis and treatment increase the risk of death and poor prognosis. The limited diagnostic and therapeutic window and the shortage of neurologists with experience in stroke management has made telemedicine a method for providing neurological care readily available. Teleneurology has been shown to be a safe, efficient, and cost-effective way of improving stroke outcomes [26]. The accuracy of telestroke in diagnosing acute stroke has proven to be equivalent to that of bedside evaluations [27].

Telestroke systems have included AI algorithms using machine learning to automate reading and classification of radiologic imaging. The efficiency and quality of care depend on the rapid interpretation of clinical data and brain images.

Identification of stroke location, classification and severity can directly impact management decisions [28].

The addition of deep learning algorithms has shown the ability to identify abnormalities such as intracranial hemorrhages and its subtypes, midline shift, and mass effect [28]. In large vessel occlusion ischemic strokes, selected patients could benefit from endovascular reperfusion therapy. The incorporation of automated perfusion processing to interpret perfusion is helping in the treatment decision-making process [29].

AI algorithms have been integrated into various processes of acute stroke management. Prehospital notifications sent to the emergency department can alert an incoming stroke, allowing for the adequate preparation of patient's appropriate management according to the type and other characteristics of the stroke (i.e., neurointerventional, neurosurgical, or neurocritical care teams) [28].

Telestroke programs are using platforms providing automated information about various components of the acute stroke triage pathway and clinical workflow. They offer fast and accurate analyses to optimize the delivery of stroke care at spoke and hub hospitals [30, 31].

Acute stroke patient candidates for thrombolysis can be rapidly detected with the help of machine learning algorithms for the identification of ischemic infarction on computerized tomography (CT) or magnetic resonance (MR) diffusion-weighted imaging. AI has been effective to identify core infarct volumes on MR DWI and large vessel occlusion, essential for selecting patients who could benefit from mechanical thrombectomy.

ASPECTS is a widely used 10-point quantitative topographic CT scan grading system for assessing the extent of early ischemic stroke in the selection process of thrombectomy candidates. ASPECTS scoring software platform (e-ASPECTS, Brainomix) performed as well as neuroradiologists in patients with acute stroke [32]. However, in patients with acute stroke with baseline non-normal-appearing CT (e.g., leukoencephalopathy, old infarcts, or other parenchymal defects), e-ASPECTS did not perform as well as neuroradiologists [33].

RapidAI is another neuroimaging software, which collects and analyses data from the CT

perfusion scan. It provides a real-time view of brain perfusion with a quantitative value of the volume of brain tissue that is irreversibly damaged and the volume of tissue that can be recovered if blood flow can be restored quickly [34].

The current application of AI in the telestroke field has allowed significant improvement in treatment decision and clinical outcomes and future studies validating AI techniques are needed to allow its widespread use.

Teledermatology and AI

Teledermatology (TD) is transforming healthcare recently. The evidence to date supports the accuracy and cost-effectiveness of teledermatology and the possibility of bringing greater access to the specialist.

Dermatological diagnosis relies mainly on morphological features, and most diagnoses are based on visual pattern recognition. Skin imaging technology has become an important tool for clinical diagnosis of skin diseases, and TD is one of the telemedicine areas in which AI image recognition capabilities for assisted diagnosis have been well explored. TD with AI integration might also be an effective screening tool. However, there are limitations as a recent systematic review of currently available teledermatology mobile-based apps concluded that studies done so far relied on heavily biased patient samples, limiting the conclusions on the efficacy of these automated-diagnoses systems [35].

In primary care setting, artificial intelligence-driven image diagnosis could help general practitioners to make a correct diagnosis by feeding a skin lesion picture to an application to diagnose common dermatological conditions. By linking such applications to a remote central enriched with image-based diagnostic database that analyzes the images, the non-dermatologist physician can receive the probabilities of diagnosis or differential diagnoses.

Several studies have evaluated AI-driven image diagnostic technology for skin cancers against dermatologist diagnostic performance, showing the same accuracy as the trained dermatologists [36, 37].

Deep learning technology stands to disrupt dermatology, both in person visits and by teledermatology.

However, studies using such technology have shown a limitation for non-skin cancer disease conditions. Low specificity and specificity for these pathologies have been found meaning that they are inaccurate enough to rely on AI-driven image analysis alone. If a certain dermatological condition is excluded from the AI model's database, a patient with this condition may be misdiagnosed with a similar-appearing condition. Bias may also be introduced when lesions that were deemed worthy of capturing via photograph or being biopsied may not be representative of the lesion type. As a result, by using images from datasets, the sensitivity of clinician diagnosis may be lower than in a normal clinic [38].

Telemedicine, Artificial Intelligence, and Education

For the responsible (ethical, safe, and quality) use of new technologies such as telemedicine and AI, it is of paramount importance to provide prime quality education and training for medical students, residents, and physicians. Despite efforts and the recent growth of telemedicine teaching in medical education institutions, there is still a great lack of education and training in this field. The knowledge and practice of telemedicine are and will be even more valuable in the near future to serve populations in need of quality medical care.

During medical school, a curriculum that integrates telemedicine training should be competency-based and outcomes-oriented with multiple teaching modalities. The latter may include (1) asynchronous lectures; (2) discussions on applications, ethics, safety, etiquette, and patient considerations; (3) faculty-supervised standardized patient telehealth encounters; and (4) hands-on diagnostic or therapeutic procedures using telehealth equipment such as live video, the store-and-forward method, remote patient monitoring, and mobile health [39].

Such curricula should allow students to learn how to maintain a strong patient-doctor

relationship, protect patient privacy, promote equity in access and treatment, seek the best possible outcomes, and learn the benefits and limitations of telemedicine [40].

The COVID-19 pandemic has caused an exponential growth worldwide in the use of telemedicine. The incorporation telemedicine and AI teaching into medical school curriculum will also increase their understanding of the complex ethical, regulatory, and legal issues related to the use of such technologies [39]. The main domains of such curriculum could include (1) access to care, (2) cost, (3) cost-effectiveness, (4) patient experience, and (5) clinician experience [39].

The education on telemedicine and AI should also be expanded for practicing physicians. The American Medical Association recommends medical specialty societies and boards to consider the production of specialty-specific educational modules related to AI [22]. AMA also highlights the need for continued medical education regarding assessment, understanding, and application of data in patients' care.

Nowadays, physicians should have skills and competencies to work with electronic health records and understand the true potential of the new technologies. They should have the ability to manage data and supervise AI applications, to use it as clinical decision support.

Physicians do not need to become experts in AI; rather they should have sufficient understanding of the capabilities and limitations of AI algorithms to allow them to make the best use of it. However, it is important to ensure that the use of AI and telemedicine does not harm the humanism enshrined in the practice of medicine and the patient-physician relationship.

References

- Pan American Health Organization. Framework for the implementation of a telemedicine service. <https://iris.paho.org/handle/10665.2/28414>. Accessed 20 Oct 2020.
- Kuziemsky C, Maeder AJ, John O, Gogia SB, Basu A, Meher S, Ito M. Role of artificial intelligence within the telehealth domain. *Yearb Med Inform.* 2019;28(01): 035–40. <https://doi.org/10.1055/s-0039-1677897>.
- WHO Group Consultation on Health Telematics. A health telematics policy in support of WHO's Health-For-All strategy for global health development: report of the WHO group consultation on health telematics, 11–16 December, Geneva, 1997. World Health Organization. <https://apps.who.int/iris/handle/10665/63857>. Accessed 24 Oct 2020.
- WHO Global Observatory for eHealth. Telemedicine: opportunities and developments in Member States: report on the second global survey on eHealth. 2009. ISBN 978-92-4-156414-4. Accessed 16 Oct 2020.
- WHO guideline Recommendations on Digital Interventions for Health System Strengthening. 2019. <https://www.ncbi.nlm.nih.gov/books/NBK541902/>. Accessed 17 Oct 2020.
- Almathami HKY, Win KT, Vlahu-Gjorgjevska E. Barriers and facilitators that influence telemedicine-based, real-time, online consultation at patients' homes: systematic literature review. *J Med Internet Res.* 2020;22(2):e16407. <https://doi.org/10.2196/16407>.
- Monaghesh E, Hajizadeh A. The role of telehealth during COVID-19 outbreak: a systematic review based on current evidence. *BMC Public Health.* 2020;20:1193. <https://doi.org/10.1186/s12889-020-09301-4>.
- Orlando JF, Beard M, Kumar S. Systematic review of patient and caregivers' satisfaction with telehealth videoconferencing as a mode of service delivery in managing patients' health. *PLoS One.* 2019;14(8):e0221848. <https://doi.org/10.1371/journal.pone.0221848>.
- Tchero H, Kangambega P, Briatte C, Brunet-Houdart-S, Retali GR, Rusch E. Clinical effectiveness of telemedicine in diabetes mellitus: a meta-analysis of 42 randomized controlled trials. *Telemed J E Health.* 2019;25(7):569–83. <https://doi.org/10.1089/tmj.2018.0128>.
- Zhu Y, Gu X, Xu C. Effectiveness of telemedicine systems for adults with heart failure: a meta-analysis of randomized controlled trials. *Heart Fail Rev.* 2020;25:231–43. <https://doi.org/10.1007/s10741-019-09801-5>.
- Hailey D, Paquin M-J, Casebeer A, Harris LE, Maciejewski O. Evidence about tele-oncology applications and associated benefits for patients and their families. *J Telemed Telecare.* 2006;12:40–3. <https://doi.org/10.1258/135763306779379941>.
- Sonu B, Sian B, Kumar CV, Anil A, Alma N, Saltanat K, et al. Telemedicine across the globe—position paper from the COVID-19 Pandemic Health System Resilience PROGRAM (REPROGRAM) International Consortium (Part 1). *Front Public Health.* 2020;8:644. <https://doi.org/10.3389/fpubh.2020.556720>.
- Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25:44–56. <https://doi.org/10.1038/s41591-018-0300-7>.
- Pacis DM, Subido ED Jr, Bugtai NT. Trends in telemedicine utilizing artificial intelligence. *AIP Conf Proc.* 2018;13:1933. <https://doi.org/10.1063/1.5023979>.

15. Faes L, et al. A clinician's guide to artificial intelligence: how to critically appraise machine learning studies. *Transl Vis Sci Technol.* 2020;9:7. <https://doi.org/10.1167/tvst.9.2.7>.
16. Anderson M, Anderson SL. How should AI be developed, validated and implemented in patient care? *AMA J Ethics.* 2019;21(2):E125–30. <https://doi.org/10.1001/amajethics.2019.125>.
17. Meskó B, Marton G. A short guide for medical professionals in the era of artificial intelligence. *NPJ Digit Med.* 2020;3:126. <https://doi.org/10.1038/s41746-020-00333-z>.
18. Downing NL, Bates DW, Longhurst CA. Physician burnout in the electronic health record era: are we ignoring the real cause? *Annals of Internal Medicine.* 2018; <https://doi.org/10.7326/M18-0139>.
19. Roshan M, Rao A. A study on relative contributions of the history, physical examination and investigations in making medical diagnosis. *J Assoc Physicians India.* 2000;48(8):771–5.
20. Bhaskar S, Bradley S, Sakhamuri S, Moguilner S, Chattu VK, Pandya S, et al. Designing futuristic telemedicine using artificial intelligence and robotics in the COVID-19 era. *Front Public Health.* 2020;8:708. <https://doi.org/10.3389/fpubh.2020.556789>.
21. Mathews SC, McShea MJ, Hanley CL, Ravitz A, Labrique AB, Cohen AB. Digital health: a path to validation. *NPJ Digit Med.* 2019;2:38. <https://doi.org/10.1038/s41746-019-0111-3>.
22. American Medical Association. Augmented intelligence in health care. 2019. <https://www.ama-assn.org/system/files/2019-08/ai-2018-board-policy-summary.pdf>. Accessed 10 Nov 2020.
23. Horton MB, Brady CJ, Cavallerano J, Abramoff M, Barker G, Chiang MF, et al. Practice guidelines for ocular telehealth-diabetic retinopathy. *Telemed J eHealth.* 2020;26:495–543.
24. Li JPO, Liu H, Ting DSJ, Jeon S, Chan RVP, Kim JE, et al. Digital technology, tele-medicine and artificial intelligence in ophthalmology: a global perspective. *Prog Retin Eye Res.* <https://doi.org/10.1016/j.preteyes.2020.100900>.
25. Fernandes JG. Stroke prevention and control in Brazil: missed opportunities. *Arq Neuropsiquiatr.* 2015; <https://doi.org/10.1590/0004-282X20150127>.
26. Powers WJ, Rabinstein AA, Ackerson T, Adeoye OM, Bambakidis NC, Becker K, et al. Guidelines for the early management of patients with acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke.* 2018;49:e46–99. <https://doi.org/10.1161/STR.0000000000001063>.
27. Agrawal K, Raman R, Ernstrom K, Claycomb RJ, Meyer DM, Hemmen TM, et al. Accuracy of stroke diagnosis in telestroke-guided tissue plasminogen activator patients. *J Stroke Cerebrovasc Dis.* 2016;25: 2942–6. <https://doi.org/10.1016/j.jstrokecerebrovasdis.2016.08.009>.
28. Ali F, Hamid U, Zaidat O, Bhatti D, Kalia JS. Role of artificial intelligence in TeleStroke: an overview. *Front Neurol.* 2020;11:559322. <https://doi.org/10.3389/fneur.2020.559322>.
29. Vagal A, Wintermark M, Nael K, Bivard A, Parsons M, Grossman AW, et al. Automated CT perfusion imaging for acute ischemic stroke. *Neurology.* 2019;93:888. <https://doi.org/10.1212/WNL.0000000000008481>.
30. Soun JE, Chow DS, Nagamine M, Takhtawala RS, Filippi CG, Yu W, Chang PD. Artificial intelligence and acute stroke imaging. *Am J Neuroradiol.* 2020; <https://doi.org/10.3174/ajnr.A6883>.
31. Albers GW, Marks MP, Kemp S, et al. Thrombectomy for stroke at 6 to 16 hours with selection by perfusion imaging. *N Engl J Med.* 2018;378:708–18.
32. Nagel S, Sinha D, Day D, et al. e-ASPECTS software is non-inferior to neuroradiologists in applying the ASPECT score to computed tomography scans of acute ischemic stroke patients. *Int J Stroke.* 2017;12: 615–22.
33. Guberina N, Dietrich U, Radbruch A, et al. Detection of early infarction signs with machine learning-based diagnosis by means of the Alberta Stroke Program Early CT score (ASPECTS) in the clinical routine. *Neuroradiology.* 2018;60:889–901.
34. Kauw F, Heit JJ, Martin BW, van Ommen F, Kappelle LJ, Velthuis BK, et al. Computed tomography perfusion data for acute ischemic stroke evaluation using rapid software: pitfalls of automated postprocessing. *J Comput Assist Tomogr.* 2020;44:75–7. <https://doi.org/10.1097/RCT.0000000000000946>.
35. Finnane A, Dallest K, Janda M, Soyer HP. Teledermatology for the diagnosis and management of skin cancer: a systematic review. *JAMA Dermatol.* 2017;153:319–27.
36. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist level classification of skin cancer with deep neural networks. *Nature.* 2017;542(7639):115–8. <https://doi.org/10.1038/nature21056>.
37. Brinker TJ, Hekler A, Enk AH, Klode J, Hauschild A, Berking C, et al. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. *Eur J Cancer.* 2019;111:148–54. <https://doi.org/10.1016/j.ejca.2019.02.005>.
38. Tschanzl P, Codella N, Akay BN, Argenziano G, Braun RP, Cabo H, et al. Comparison of the accuracy of human readers versus machine learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *Lancet Oncol.* 2019;20(7):938–47. [https://doi.org/10.1016/S1470-2045\(19\)30333-X](https://doi.org/10.1016/S1470-2045(19)30333-X).
39. National Quality Forum. Creating a framework to support measure development for telehealth. https://www.qualityforum.org/Publications/2017/08/Creating_a_Framework_to_Support_Measure_Development_for_Telehealth.aspx. Accessed 28 Nov 2020.
40. Jumreornvong O, Yang E, Race J, Appel J. Telemedicine and medical education in the age of COVID-19. *Acad Med.* 2020; <https://doi.org/10.1097/ACM.0000000000003711>.



AIM and mHealth, Smartphones and Apps

88

Joseph Davids and Hutan Ashrafiyan

Contents

Introduction	1230
History	1230
AI and mHealth in Various Medical Specialties	1235
AI and mHealth for Evidence-Based Medicine	1235
AI and mHealth in the Field of Genomics	1236
AI and mHealth in the Field of Cardiovascular Medicine	1236
AI and mHealth in the Field of Respiratory Medicine	1237
AI for mHealth for Neuroscience and Neuropsychiatric Disorders	1238
AI in mHealth for Rheumatology	1238
AI in mHealth for Gastroenterology	1239
AI in mHealth for Urology	1239
AI in mHealth for Endocrinology	1239
AI in mHealth for Dermatology	1240
AI in mHealth for Obstetrics and Gynecology and Pediatrics	1241
AI in mHealth for Consensus Evaluation	1241

J. Davids ()

Imperial College London NHS Trust, London, UK

Hamlyn Centre for Robotics and Artificial Intelligence,
Department of Surgery and Cancer, Imperial College
London, London, UK

National Hospital for Neurology and Neurosurgery Queen
Square, London, UK
e-mail: jdavids@ic.ac.uk

H. Ashrafiyan
Department of Surgery and Cancer, Imperial College
London NHS Trust, London, UK

Institute of Global Health Innovation, Imperial College
London, London, UK

Hamlyn Centre for Robotics and Artificial Intelligence,
Department of Surgery and Cancer, Imperial College
London, London, UK
e-mail: h.ashrafiyan@ic.ac.uk

AI and mHealth in Infectious Diseases	1242
Summary and the Future of mHealth	1242
References	1243

Abstract

Incremental advances in miniaturized transistor technologies and improvements in network infrastructure have paved the way for the development of smartphones and applications augmenting the management of various conditions ranging from respiratory to neuropsychiatric disorders. Other applications have also been borne out of the necessity to streamline services in order to tackle logistical challenges that often arise in medicine and healthcare. These include using smartphone apps to aid diagnostics, such as skin lesion classification, or to leverage online platforms and facilitate easier access to patient information. What would otherwise take up additional unnecessary hospital resources is now achievable through mHealth systems such as patient flow monitoring. The ease with which information can now be accessed, curated, and pre-processed has paved the way for the concept of medical informatics. Informaticians now have the capability to apply artificial intelligence and machine learning to derive critical insights, automate, and make predictive analysis to improve clinical and quasi-clinical healthcare delivery. This chapter explores the global evolution of the artificial intelligence paradigm for mHealth, eHealth, and smartphone and mobile phone apps for medicine. The chapter outlines some of their applications for various medical subspecialties including cardiorespiratory, neuropsychiatric, and rehabilitation medicine.

Keywords

mHealth · Artificial intelligence · Smartphone apps · Machine learning · Deep learning · Supervised learning

Introduction

The fields of telehealth and mHealth are defined by the use of smartphone, network, and wireless technologies to provide direct or indirect health support to patients. They also facilitate technical support for healthcare providers and various healthcare organizations [1]. This has not only received a lot of attention from the global health community, but has also resulted in a considerable amount of innovation to streamline medical care, interventions, patient flow, and hospital logistics coordination – all leading to improvements in multidisciplinary care. These developments were made possible thanks to considerable nanotechnological advances in top-down miniaturization of transistors and other semiconductor chips to create devices that allow for better software-to-hardware integration [2], thus enabling healthcare engineers to facilitate the design and platform execution of multi-agent-based artificially intelligent algorithms. This in turn led to the conception, generation, and deployment of structured machine learning pipelines that can be optimized to work on a mobile device for the benefit of both the patient and the healthcare system. To understand how AI impacts and augments mHealth, we explore some of the key historical milestones that brought us to where we are today. Since earlier chapters have already outlined the origins of AI and machine learning, this will not be a focus in this chapter. However, we will briefly align the chapter's agenda with the origins of networks that paved the way for the mHealth, eHealth, and smartphone app revolution over recent decades.

History

First-generation (1G) mobile networks started with the Japanese Nippon Telegraph and Telephone company in the late 1970s, evolving from

pre-wireless 0G technologies that developed after World War II. The mobile phone industry has made remarkable incremental strides since then, reaching its fifth generation of innovation status in recent years. Many initial drawbacks included the cost, which at the advent of 1G was \$3,995 (present US inflation adjusted cost of \$26,648.19; a 567.0% rise), unreliably poor voice linking, low handoff, and poor network security, which subsequently improved for both the developed and developing worlds. Although 1G was remarkably cost-prohibitive, it still saw widespread transatlantic global adoption with Motorola's DynaTAC being the first to deploy it in the USA in 1983 as well as developments in Europe with popular analogue systems from Nordic Mobile Telephone (NMT) and TACS [3–5].

Second-generation (2G) mobile networks started a cultural revolution and was built on the Finnish contribution of the Global System for Mobile Communications (GSM) standard. 2G not only added a layer of security to wireless telecommunications through encryption systems, but also added clearer digital voice calling with less static interference, thus making it more usable for robust machine learning applications on the audio itself. The introduction of picture and multimedia messages then extended its mass consumer adoption, paving the way for innovations in computer vision applications for mHealth. The 2G era also saw transfer speed innovations from 10kbs to 500 kbps at its bleeding edge [3–6].

In 2001, the packet switching third-generation (3G) mobile networks not only elevated network technology but also introduced standardizations that made these technologies ubiquitously adoptable and easily accessible from all corners of the globe. Launched by NTT DoCoMo, roaming became available for the first time from any international location with data packet transfer-driven technology [5]. With transfer speeds of up to almost 2 Mbps, four times the speed of its predecessor, such augmentations enabled quality video conferencing, streaming, and voice-over Internet protocols (adopted by companies like Skype) [3–6]. Derivations of 3G included EDGE, CDMA2000 [6]. Readers may recall the era of the personal digital assistant, such as the BlackBerry,

with mobile platforms that were highly coveted at that time due to their impressive professional capabilities, technological fidelity, and portable form factor, that is, until they were superseded by the Apple iPhone (see Fig. 1).

4G kick-started the so-called 100Mbps streaming era and introduced the Long-Term Evolution standard, which made high-quality video streaming a reality on small form-factor devices, at the same time parallelizing compute resources to minimize bandwidth deficiencies that the previous generations all harbored [6]. Scandinavian countries such as Norway and Sweden were early adopters of the technology, followed by hundreds of millions of global consumers, thus paving the way for this technology to be utilized by patients, family members, and healthcare teams around the world [3–5]. 4G formed the bedrock for the development of mHealth-powered AI in various medicine-related settings, including augmented reality, medical games, high-definition video conferencing, mixed reality streaming, and other methods utilized for medical ward rounds by centers like Oxford University and Imperial College Hospitals NHS Trust in the UK [7, 8]. Figure 1 summarizes the evolution of some of the smartphone and wearable technologies that are being used in mHealth. This subsequently became crucial during the 2019 SARS-CoV-2 pandemic allowing for adherence to social distancing rules, keeping the quintessential doctor-patient relationship alive while still maintaining effective healthcare delivery and patient safety.

The interlink between 5G and the evolution of the Internet to support it requires a joint treatise. The concept of the Internet itself was borne out of the idea of wireless machine telegraphy circa 1830s [9, 10]. James Clerk Maxwell's theory of electromagnetic radiation was proposed in a lecture to the Royal Society of London in the mid- to late nineteenth century [9]. More than 20 years later, in 1887, Heinrich Hertz in Germany confirmed Maxwell's theories. Irish-Italian wireless pioneer Guglielmo Marconi was the first to equip the Titanic with commercially available wireless telegraphy during its maiden voyage in 1912 [9]. Coupled with the advent of radio-voice communication at the turn of the twentieth century allowed the origins of the Internet and its



Fig. 1 A historical overview of mobile technology platforms. From left to right, up to down. Digital artistic impression of Guglielmo Marconi with a wireless telegraphy system c. 1902, adapted from the Science Museum Collection, Wireless Telegraph and Signal Co. Ltd. Top row middle column image, the Motorola DynaTAC by Mike Kuniavsky from the science museum. Top row right middle column, first BlackBerry Series image by CrackBerry. Top row upper far right is the first-generation

iPhone photo by Rafael Fernandez; 5G-capable Samsung smartphone from Samsung. Top row, lower middle is the Apple watch series 6. Top row bottom right, Focals by North image from Wired. Bottom row the HoloLens 2 digital sketch adapted from Microsoft Image [12]. (Disclaimer: This is an independent publication and is neither affiliated with nor authorized, sponsored, or approved by Microsoft Corporation)

protocol-driven technologies such as the Internet of Things (IoT) to come to the fore. However, it was not until the advent of early computers in the 1950s at the height of the Cold War that led the Defense Advanced Research Projects Agency (DARPA) to create the ARPANET, conceptualized by JCR Licklider and created in the late 1960s as the predecessor to the World Wide Web

that we know today as pioneered by Sir Tim Berners-Lee and others [10, 11]. At the same time, Global Position Satellites (GPS) developed during the late twentieth century became another primary component for IoT, together with landlines [3–5, 11].

At the time of writing, 5G has not been widely adopted because of marked controversies and

geopolitical tensions; however the concept of the Internet of Things (IoT) and connectivity is not new. Prof. Kevin Ashton is credited with coining the term in the 1990s to secure funding for Radio-frequency Identification tagging Technology (RFID) from Proctor and Gamble [10]. This technology has allowed devices to be tagged and tracked over appreciably long distances. Other complementary short-range methods like QR and barcode technology for digital watermarking have also facilitated digital inventory control that is useful for device-led machine learning and AI approaches in medicine.

An insightful speech by Prof. Ashton outlined the potential of IoT for medical applications but also highlighted the challenges we face as human beings and clinicians when he addressed the management of big data and the need for a paradigm shift in approach. In 1999, he remarked that [10]:

...nearly all of the roughly 50 petabytes (a petabyte is 1,024 terabytes) of data available on the Internet were first captured and created by human beings by typing, pressing a record button, taking a digital picture or scanning a barcode. The problem is, people have limited time, attention, and accuracy. All of which means they are not very good at capturing data about things in the real world. If we had computers that knew everything there was to know about things, using data they gathered without any help from us, we would be able to track and count everything and greatly reduce waste, loss and cost. We would know when things needed replacing, repairing or recalling and whether they were fresh or past their best. [10]

Ashton also postulated the tagging of devices with RFID to provide tracking potential. Early research development by NASA's machine to machine intelligence corps and South Korea's 5G research and development group contributed to where we are today. In fact, Korean companies like KT, LG Uplus, and SK Telecom had already started providing commercial 5G services as early as the end of 2018. The speed of 5G is unparalleled compared to its predecessors with an estimated upload and download speed increase over 20 times the previous generations (up to 20 Gbps) [6–8, 11]. Figure 2 summarizes the global network subscribership over the past 27 years and the most recent data on the number of Internet users.

We now have pervasive computing devices including the latest generation of Apple/Samsung/Fitbit, etc. smart watches/wearable devices that can be used for medical monitoring. For example, Apple watch device can network with smartphone apps using machine learning algorithms to detect cardiac electrophysiological abnormalities and pulse abnormalities. Upon detection, they are able to send the results wirelessly over a 4G network to a specialist cardiologist or general/family practitioner's phone, thus facilitating the early detection of potentially serious conditions. As this example shows, we have already come a long way with AI, but there is still so much more that can and will be accomplished.

The International Telecommunication Union reports 5 billion users of wireless technologies, and a recent survey of 114 countries conducted by the WHO Global Health Observatory categorized about 16 facets of mHealth services as follows [15]:

- Mobile telemedicine
- Raising health awareness
- Medical decision support systems
- Medical treatment compliance
- Managing medical emergencies
- Managing medical disasters
- Health call centers
- Emergency toll-free telephone services
- Medical appointment reminders
- Community mobilization and health promotion
- Mobile patient records
- Information access
- Patient monitoring
- Health surveys and data collection
- Medical surveillance
- Raising health awareness

The four most common areas of mHealth across the globe include health call centers (59%), emergency toll-free telephone services (55%), managing emergencies and disasters (54%), and mobile telemedicine (49%). This is in line with the historical and technological developments pioneered to facilitate mHealth [15]. mHealth now has the potential to enable disseminated cross-border healthcare delivery for

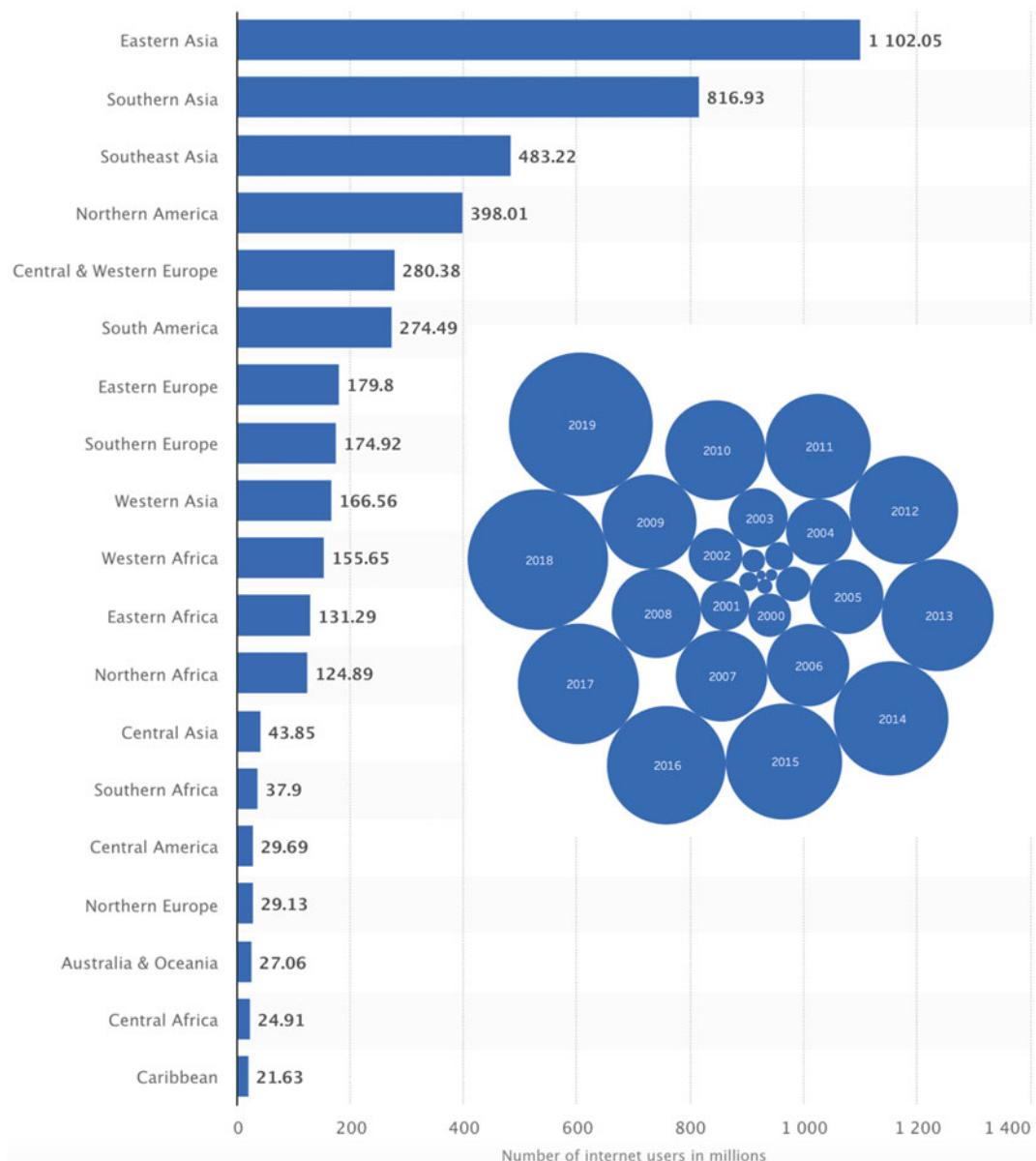


Fig. 2 Graphed data on the number of worldwide Internet users in 2020, by region (in millions), together with a yearly distribution of global mobile subscribership in

millions indicating a growing trend – the highest recorded was in 2019. (Data from [statistica.com](#), [world bank](#) [13, 14])

the benefit of low- to middle-income countries. Multiple governments have expressed their interest in helping to strengthen health systems in line with the Millennium Development Goals [15]:

“... R.D Gardi Medical College carried out a year-long research project that focused on how the use of mobile can benefit reporting on childbirth within

the province. The research covered the population of 1.9 million in the entire district of Ujjain in India.” He continued: “Through the use of mobile, we were able to collect information from 1,800 village works or community members from 152 villages. We relied on them to report their own information, but the use of mobile technology meant that our researchers were able to follow up and verify this information the next day.”

Vishal employed a call center and invited community members to call in to report a new childbirth event. This had a 68% response rate and also revealed that 20,000 prescriptions were taken over 6 months from 9 districts. Vishal added: "There are over 500 unqualified medical providers working in the province. The mobile technology meant we were able to contact the majority of these practitioners to make use of their reporting on child births in the province." Vishal is of the opinion that using mobile technology to conduct research in low- or middle-income countries is highly beneficial and will continue to use this form of reporting.

Above is an intriguing example where mHealth was used to make an impact in this rural India case study reported in the *British Medical Journal* [16, 17].

AI and mHealth in Various Medical Specialties

This section explores the use of mHealth and smartphone applications in various medical specialties, highlighting where they have been used to facilitate patient and clinician education as well as the effective diagnosis and/or treatment of various conditions. We present some of the evidence, where available, for the use of AI and suggest where other artificial intelligence systems could be effective in augmenting mHealth solutions presented in these areas. While the majority have been focused on model development, we suggest, where possible, how some of these AI agent-based models can be extended into mobile apps to support mHealth.

AI and mHealth for Evidence-Based Medicine

The scientific and clinical community are always looking for high-quality, evidenced-based research methods leading to high degrees of evidence-based care. Hierarchically, this includes:

- Systematic reviews and meta-analyses of randomized controlled trials (RCT)
- RCTs with definitive results with non-overlapping confidence intervals

- RCTs with non-definitive results and a clinically significant effect, but with confidence intervals overlapping the threshold for this effect
- Cohort studies
- Case-control studies
- Cross-sectional surveys
- Case reports

Figure 3 illustrates applications of mHealth and AI at various stages of the evidence-based pyramid for the conduct of systematic reviews and meta-analysis using platforms that enable the conduct of systematic reviews.

All of these evidenced-based approaches, but especially clinical trials and systematic reviews, tend to be both extremely costly and time-consuming to conduct, which remain some of the greatest obstacles for the conventional research methodology. Twenty-first-century developments support a clear need to employ novel mHealth systems to address research questions through judicious use of artificial intelligence. This is especially true for analysis purposes. There are tools in existence that allow algorithmic keyword PRISMA-guided search methodologies on various databases [18, 19]. The ability for ubiquitous search term standardization across multiple databases developed by Bond University in Australia known as the S-R Accelerator should be performed simultaneously [18, 19]. Other mHealth apps like Rayan now allow for systematic cross-collaborative screening of the searched data and enable consensus to be reached [20]. Other areas for AI augmentation also include randomization methods for randomized controlled trials, which can now be conducted via online systems such as Criteria2Query that leverage natural language processing [21].

On the conduct of systematic reviews on mHealth, a large meta-analysis of mHealth for hospital appointments reported that in 42 trials, there was significant risk of bias in the conduct of trials, but modest benefits for mHealth in healthcare provider intervention support and management outcomes [22].

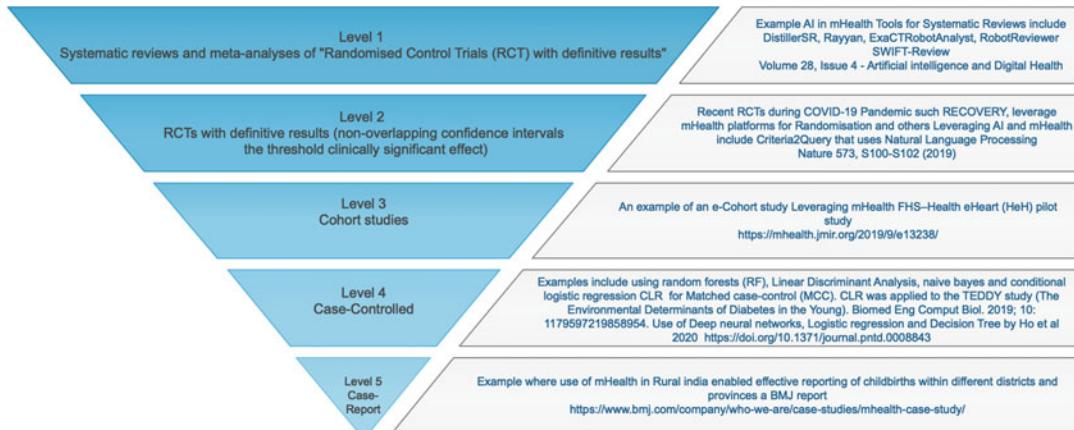


Fig. 3 Evidence-based medicine pyramid. Inverted here to illustrate the weight that each section carries. Examples are provided of how mHealth has made an impact and been used at each of these levels

AI and mHealth in the Field of Genomics

Genomics studies the function of encoded genetic information in the DNA sequences of all organisms. It has seen an explosion of growth in recent years since the entire human genome, containing 20,000 protein coding and 25,000 non-protein coding genes was sequenced in 2001 [23]. Current mHealth platforms enable the identification of specific genetic and epigenetic risk factors, such as copy number variants, to help stratify risks of psychiatric and other diseases. These platforms include the National Cancer Institute's Cancer Genome Atlas project, which has not only developed algorithms for the visualization of large genomic bioinformatic datasets including proteomics and radiomics, but has also developed portals to facilitate data access via mobile devices and tablets [23]. Available structured and labeled datasets for risk stratification and outcome analysis are key and have opened up a whole new world of application programming interface development for mobile phone apps.

Other noteworthy reviews by Leung and colleagues offer a comprehensive overview of machine learning in the field of genomics [24]. They outline that while we have a comprehensive instruction book for the coded make-up of humans, the blueprint of how our genes are organized still remains encoded, which is challenging

in terms of understanding various diseases. Here is where artificial intelligence shines in terms of pattern recognition, latent variable identification, distribution analysis and feature engineering, etc.

Genomics is now allowing ancestral linkage mapping on a global scale, with companies like 23 and Me building business models to capitalize on the genetic mHealth revolution and personalised medicine. The EU General Data Protection Regulation and other ethico-legal considerations and ramifications rightfully limit the extent of how these apps can be used, but they can be developed further under closer regulatory scrutiny to prevent abuse.

AI and mHealth in the Field of Cardiovascular Medicine

In cardiovascular medicine, patients recovering from myocardial infarction require cardiac rehabilitation (CR). CR remains an important evidence-based intervention allowing effective progression to patient recovery following an acute coronary event. Patients receive a structured education and exercise program, which helps to accelerate their recovery time. It also allows lifestyle and behavioral recalibration to reduce the risk of further deterioration in cardiac health. The NHS has approved an mHealth app called the myHeart app, which enables structured

educational and exercise intervention. In one study, 721 participants were placed in one of four groups (class-based CR, class-based CR with mOverallyHeart, home-based CR, and home-based CR with myHeart) [25]. Five hundred and eighty-four patients elected to use class-based CR. Of these, 43 chose to include myHeart to support their rehabilitation [25]. Artificial intelligence can also be applied to support patients looking to augment their lifestyle choices. Recommendation systems constructed from deep neural networks can be integrated to suggest healthy lifestyle alternatives and to widen the choice of food, exercise regimens, and personalized health plans for the patients given the overwhelming amount of available choice [26–29]. In order to achieve a holistic approach, clinicians, dieticians, occupational health specialists, and physiotherapists could be invited to facilitate expert, directed lifestyle interventions on these types of eHealth platforms without the need for face-to-face contact [30].

This could help reduce the patient's risk of recurrent coronary events, reduce their cholesterol levels, and facilitate smoking cessation program e-support, if needed. Chatbots can automate the process, and their integration can facilitate the dissemination of these recommendations to a particular demographic who are at risk of relapse into previous unhealthy habits [31–47].

Another interesting development relates to the release of Apple's iOS 12 with the Apple HealthKit. Their closed ecosystem has some excellent advantages for privacy protection but also allows individuals and medical application developers to build secure permission-enabling apps such as those for heart rhythm analysis. In fact, at the launch, Apple showcased case-studies of individuals who were immediately identified as having acute coronary events and who went on to receive the emergency clinical intervention needed to save their lives. Others reported that AI algorithm-based smartwatch technologies worked well compared to an implantable loop recorder for the detection of atrial fibrillation and will therefore become useful for population mHealth screening [48, 49].

AI and mHealth in the Field of Respiratory Medicine

Chronic obstructive pulmonary disease (COPD) is a common preventable condition and a leading cause of death and disability-adjusted life years. It has considerable physical and psychosocial morbidity affecting around 80,000 people in the UK alone and is a leading cause of hospital admissions [50]. mHealth has also seen significant applications in this field, including educating the public through smoking cessation programs and pulmonary rehabilitation. MyCOPD and others are example of online application support mHealth platforms for patients with COPD receiving education on self-monitoring and self-management [50–52]. MyCOPD offers the capability of implementing and assessing how-to instructional videos with the potential for artificial intelligence chatbot integration to support patients. In their non-inferiority randomized controlled trial of patients aged 40–80 with mild to moderate COPD, comparative outcomes were obtained for online mHealth methods of pulmonary rehabilitation (myCOPD group) against in-person pulmonary rehabilitation. Moreover, the odds of ≥ 1 critical inhaler error was marginally significantly lower in the myCOPD arm (adjusted odds ratio (OR) 0.30 (95% CI 0.09–1.06, $p = 0.061$)). Secondarily, a noticeable increase in the app's usage was linked with a greater CAT score (which measures the recovery rate of symptoms) improvement. The adjusted odds ratio for being in a higher patient activation measurement (PAM) level at 90 days was 1.65 (95% CI 0.46–5.85) in favor of myCOPD. The RESCUE trial also looked at application supported care vs usual standard of care and found superior outcomes with less exacerbations and readmissions than the non-mHealth arms [OR 0.383 (95% CI: 0.074, 1.987; $n = 35$)] [52].

At the time of writing, the 2019 SARS-CoV-2 coronavirus pandemic had curtailed the delivery of face-to-face services with considerable impact on high-risk individuals with COPD. The advantages of mHealth can be noted in its protective role in helping limit and tracking the spread of viruses. mHealth platforms can allow multiple

appointments to be delivered without additional problems with scheduling. Classes could be delivered at the same time to a large number of individuals, whereas in-person scheduling conflicts can affect the delivery of face-to-face sessions. Logistically, a reduction in readmission rates has also been seen with mHealth supporting respiratory conditions such as COPD and asthma [50–52]. Other examples where mHealth was making an impact include Stanford University's 100,000 dataset trained 121-layer convolutional neural network called CheXNet, where an AI platform can diagnose pneumonia based on patient chest radiographs [53]. Other areas such as screening for obstructive sleep apnea have also been studied using artificial intelligence, and apps have also been demonstrated [54].

AI for mHealth for Neuroscience and Neuropsychiatric Disorders

The concept of connectomics looks at the connectivity within the brain from clinical, genomic, physiological, electrophysiological, biochemical, and computational perspectives, uniting these datasets to allow clinicians to manage complex conditions. The Human Connectome and Human Brain projects have utilized platforms to integrate and provide an unparalleled compilation of neural data [55, 56]. The use of AI and mHealth to create an interface to graphically navigate this large dataset providing the tools to derive actionable insights from within the data [56]. It is believed that doing so will facilitate a holistic and semiotic approach to understanding the manifestations and progression of neurological and psychiatric diseases, such as Parkinson's disease, depression, epilepsy, etc.

One area of mHealth with AI augmentation for neuroscience in which machine learning algorithms seem to have been effective was the use of regular expressions in natural language processing to study important risk factors for sudden unexpected death in epilepsy (SUDEP), an important cause of mortality in epilepsy. Barbour et al. identified a gap in how often providers counsel patients about SUDEP and offered the potential

solution of electronically prompting clinicians to provide counseling via automated detection of risk factors in electronic medical records (EMRs) [57]. Apps can be created to support the clinician and in particular the patient by prompting them to be aware of risk factors that could increase the likelihood of SUDEP. An extended review of AI in neuro-electrophysiology is provided with practical aspects in that chapter.

Another example is the smart crisis cross-national comparative study set up to determine whether suicide risks were partly related to sleep quality and poor appetite [58]. This study was conducted by French (University hospital of Nimes) and Spanish (Hospital Fundación Jiménez Díaz Psychiatry Department) institutions in their outpatient population demonstrating the globally diverse use cases for AI and mHealth.

AI in mHealth for Rheumatology

Recently, AI has been receiving increasing interest in the field of rheumatology. Algorithms developed by Versus Arthritis, a UK Charity, and provided to IBM Watson are being used to build an mHealth platform that will support patients to better self-manage their condition [59, 60]. For this mHealth platform, IBM Watson used a collection of algorithms and systems from Versus Arthritis that could learn from over 15,000 pages of unstructured data to develop a chatbot available 24 h a day that can interact with patients from multiple countries who are then directed to relevant and reliable information about rheumatological diseases such as rheumatoid arthritis, etc. Other applications such as symptomate.com utilize chatbots for e-diagnosis [61], whereas the Isabel symptom checker works as a triaging tool and was trained on over 6000 cases [62].

Shiezadeh and colleagues also gathered data from 2500 patients attending an Iranian clinic and used ensemble machine learning and support vector machines to build a rheumatoid arthritis (RA) detector, from which they extracted the 11 most impactful features [63]. They identified these as painful elbow and knee joints, sex, number of affected joints, and erythrocyte sedimentation rate

(ESR) test results. Their best model yielded 85% accuracy and sensitivity/specificity of 44%/74% respectively. A large, highly sensitive, and accurate (92% accuracy, with sensitivity/specificity of 86%/94% respectively) UK-based study with a training set of >15,000 and testing set of >5000 patients from two clinics in Wales used a random forest model to identify eight treatment-related predictors for disease-modifying anti-rheumatoid drugs, prednisolone or methotrexate [64]. Another large-scale early-disease risk assessment study by Chin and colleagues from Taiwan that intended to discover hidden factors enabling clinicians to formalize RA diagnosis was conducted on a dataset of 1000 RA and 500,000 non-RA patients. It used matrix factorization to identify latent risk factors and trained these on a support vector machine in order to identify early-stage RA with sensitivity and specificity of ~74 and ~70%, respectively [65]. Other studies have been done on seronegative arthritides, such as systemic lupus erythematosus, and the prediction of disease progression with mHealth components [66, 67]. The use of mHealth here can extend the identification of at-risk patients with these impactful features and help to optimize therapies for early diagnosis and treatment.

AI in mHealth for Gastroenterology

A recent review showed that about 53 studies have used AI to detect malignant and premalignant intestinal lesions with models ranging from support vector machines, convolutional neural networks, regression, gradient boosting, and Gaussian mixture models [68]. They showed that most ($n = 48$) were focused on endoscopy with a small number extracting features that were mainly demographic, cardiovascular comorbidities, concomitant medication, and digestive symptoms. Another 27 studies were dedicated to improving diagnostic accuracy in cases of colorectal polyps or cancer [68].

The extremely high mortality linked with pancreatic adenocarcinoma means that early detection is paramount for surgery to improve outcomes. Twenty-two studies assessed an AI's potential to help patient identification of liver

and pancreatobiliary disorders [68]. Various studies also tested AI in the detection of pancreatic adenocarcinoma, based on endoscopic ultrasound or markers in serum samples [68, 69]. It was reassuring in some of these studies that patients with pancreatic cancer were identified to a significant degree of accuracy, with an area under the receiver operator characteristics curve of approximately 90%. Future mHealth platforms constructed around the screening of biomarkers and radiomic features can support screening programs for pancreatic conditions.

AI in mHealth for Urology

A PRISMA-guided, Prospero-registered systematic review conducted between 1994 and 2018 reported AI applications in urology including its utilization on ultrasound data for radiomic feature identification to automate cancer detection [70]. It was used to improve outcome prediction and digitize pathological tissue specimen images and for the connectomic combination of patient clinical data with biomarkers gene expression to assist disease diagnosis or outcome prediction. Some studies also applied machine learning methods to plan brachytherapy and radiation treatments, while others used video-based or robotic automated performance metrics to objectively evaluate surgical skill [70, 71]. Compared to conventional statistical analysis, 71.8% of studies concluded that AI is superior in diagnosis and outcome prediction [70]. mHealth extensions include platforms that accelerate the histological diagnosis for the surgeon. In fact, mHealth augmented reality and AI platforms are being leveraged to support the surgeon in robotic procedures to identify locations where tumors and important blood vessels are at risk to improve patient safety during kidney tumor resections [72].

AI in mHealth for Endocrinology

One major international study on patients with diabetes collated a 370 country-year survey with

data on 2.7 million individuals worldwide since 1980 found that the total number of people with diabetes in 2008 was 347 million, more than double the number in 1980 [73]. It also reported that the prevalence of diabetes has risen, or at best remained unchanged, in virtually every part of the world over the last three decades. Type 2 diabetes mellitus (T2DM) remains a global epidemiological challenge affecting both middle- and high-income economies [74]. In a multi-center service evaluation of 83 T2DM patients recruited to use an mHealth app monitored over a 12-week period, 28 chose to use myDiabetes alone, 35 chose only usual care, and 20 chose to use both [75]. During this evaluation, changes in diabetes-related clinical health outcomes such as the patient's HbA1c, blood pressure, and body mass index were monitored. Completed questionnaires highlighting problem areas in diabetes (PAID) were used to evaluate improvement markers in diabetes-related distress. Results showed that the mHealth app was acceptable in this care setting with 31 of 42 patients using it alone or as an adjunct to usual care. Patients using this mHealth app showed the greatest improvement in HbA1c (-7.5 vs -4.4 mmol/mol), systolic blood pressure (-12.2 vs $+3.3$ mmHg), and PAID score (-6.8 vs -5.2). Five hundred and eighty-six educational videos on the app platform were consumed, which accounted for each patient watching 22.5 (SD 19.6) videos. This was reflected in the reduction in PAID scores across all arms, with the app-only arm showing the greatest improvement. Here such an app with the permission could use AI to monitor usage and alert the clinician about diabetes-related distress. Transfer learning of models trained on facial emotions could be leveraged to detect distress in elderly or disabled patients using the app who permit the use of their camera on the app to help detect and alert the clinician or family member to distress.

In the study of thyroid diseases, the cytological features and morphometric analysis sometimes present a challenge to training diagnosticians learning about cytopathology. In this regard, neural networks have also been applied to help improve the diagnosis of fine needle aspiration cytological (FNAC) histological smears. In fact, Savala et al. (2018) applied machine learning to

the morphometric analysis of FNAC [76]. The FNAC smears of histologically proven cases of follicular adenoma ($n = 26$) and follicular carcinoma ($n = 31$) patients were used with a train test ratio of 39:9. The cytological features were analyzed semi-quantitatively by two independent observers, and an ANN model was built to differentiate follicular adenoma from follicular carcinoma. Performance of the ANN model was assessed by analyzing the confusion matrix and receiver operator characteristic curve. Their study reported a 100% diagnostic accuracy or ROC of 1. This system can be incorporated into an mHealth mobile phone application to further support trainees learning about how to diagnose and effectively differentiate two different types of lesions.

AI in mHealth for Dermatology

Esteva and others have proved that the capability of convolutional neural network (CNN) performance was comparable if not superior to most dermatologists in differentiating benign from malignant lesions [77]. Their CNN was trained on a dataset of 129,450 clinical images (including 3374 dermoscopic images); performance was compared to that of 21 board-certified dermatologists regarding melanoma and keratinocytic neoplasm classification using clinical images and melanoma classification using dermoscopic images. In another study by Brinker et al., dermatologists correctly classified skin lesions up to 67.2% (95% confidence interval [CI]: 62.6%–71.7%) and 62.2% (95% CI: 57.6%–66.9%). By contrast, the trained CNN achieved a higher sensitivity of 82.3% (95% CI: 78.3%–85.7%) and a higher specificity of 77.9% (95% CI: 73.8%–81.8%) [78].

Esteva et al. rightly mention an extension for mDiagnostics in the form of a mobile phone app that could facilitate self-diagnostics of a lesion outside the clinic, and this can be achieved by leveraging the smartphone camera. mHealth applications have also been designed to teach the dermatology trainee the latent features within the image that differentiates a benign from a malignant mole (see chapter "AI and Medical Education").

AI in mHealth for Obstetrics and Gynecology and Pediatrics

Obstetrics and gynecology is a highly litigious speciality with the NHS reporting in 2017 and 2018 a sum of £4513.2 million in total indemnity payments (NHS). Forty-eight percent were obstetric-related and 2% were gynecology-related negligence claims, which accounts for 15% of all indemnity payments in NHS litigation [79]. AI and mHealth decision support platforms are useful in terms of improving awareness to patients and clinicians and for risk factor identification. This can help to mitigate the number of litigations.

There is also a considerable psychosocial burden linked to errors in O&G specialties with long-term socioeconomic consequences of adverse events such as hypoxic brain injury and shoulder dystocia-related Erb paralysis, which can lead to significant disability. This is an area of active research investigating algorithms that could support obstetric clinicians with the prediction of conditions like shoulder dystocia [80].

Two studies are discussed here about cardiotocographic monitoring of fetal state [81, 82]. In one study of 2126 recordings, an e-Health platform was built for automated cardiotocographic dataset analysis which was able to differentiate normal and abnormal CTG patterns with a sensitivity of 100% and specificity of 99%. This is the SisPorto 2.0 system which was validated on over 6000 pregnancies [81]. The second leveraged support vector machines, artificial neural networks, convolutional neural networks, recurrent neural networks, and random forests in their machine learning pipeline and reported the MKNet-convolutional neural network to have the most optimal prediction results but difficult to implement [82]. Early warning decision support system application could be designed to warn the obstetrician of a predicted pathological change sequence in the fetal heart rate [81, 82].

Evidence-based personalized pediatric medicine using mHealth such as automated televillages for child health including childhood immunization coverages and in identifying pediatric respiratory disease have been reported [33, 83, 84]. Other applications for mHealth could include non-accidental injury detection mHealth platforms,

which would be of benefit, but run the risk of unintended bias.

AI in mHealth for Consensus Evaluation

Effective training requires communication skills, carefully designed assessments, and the opportunity for purposeful and useful feedback to help safeguard patients against iatrogenic-related and human factor-related error. Dias et al. and Chan et al. have both reviewed and reported on the use of machine learning to assess clinician competency and the challenges associated with implementing this approach [85, 86]. Objective structured clinical examination enables individuals to be assessed on the practical aspects of medicine after they have received training in a specific domain of medical education. The process has always required the presence of a body of responsible individuals who are experts in the field of medical assessments and a standardized approach to the questions being asked, both of which make this a time-consuming and expensive process. One particular disadvantage is that assessors usually have inherent unintentional subjective biases as to what and how they expect performance [87].

Ryan et al. and Butow and Hoque first proposed using machine learning (ML) to improve the evaluation of communication skills [88, 89]. Jani et al. built on this work to use machine learning methods that have now been applied to extract communication and history-taking skills in OSCE transcripts to reduce biased assessment [90]. For example, in their study, Jani and colleagues collated one hundred and twenty-one transcripts of two OSCE clinical scenarios (weight loss [scenario 1] and abdominal pain [scenario 2]) manually annotating each utterance across 19 communication skills and content areas. Combining two types of natural language processing algorithms, Bag of Words and GenSen, utterances were converted to two types of numeric sentence vector representations and trained on a support vector machine, conditional random field, and a bidirectional long short-term memory model. First, these ML models (MLMs) were evaluated using a five K-fold cross-validation technique on all transcripts in a weight loss

scenario to generate precision and recall and their harmonic mean, F1 scores. Second, ML models were trained on all 101 transcripts from scenario one and tested for transferability on 20 scenario two transcripts. A bidirectional LSTM was utilized. Performance analysis with K-fold cross-validation showed high mean F1 scores: median 0.87 across all 19 labels. However, this was significantly higher using a bidirectional LSTM, demonstrating significantly higher performance and transferability score ($P < 0.005$) [90].

Historically, the Delphi method is a process used to arrive at a group opinion or decision by surveying a panel of experts. The experts' responses to several rounds of questionnaires are aggregated and shared with the group after each round. Adjustments are then made to the answer each round, based on how they interpret the "group response" provided to them. The *consensio ultimum* reflects what the group actually agrees. Questionnaires can now be disseminated and results automated using mHealth apps designed for this task.

AI and mHealth in Infectious Diseases

Diseases that are known to impact maternal and child health can benefit from AI augmented mHealth, which is already being applied in maternal and child health. Current programs aim to reduce poverty-linked infectious and transmissible diseases such as HIV/AIDS, malaria, and tuberculosis (TB) in lower- to middle-income countries [91]. Developments by Microsoft using mHealth and AI to track the movement of mosquitoes have also combined robotics and artificial intelligence to manage malaria and dengue fever in tropical endemic regions with these conditions [91]. During the 2019/2020 SARS-CoV-2 viral pandemic, mHealth continued to evolve to become one of the most reliable forms of socially distanced methods of eApps to detect and track infected individuals, e-Healthcare through eClinics, eWardrounds, and other teleHealth methodologies, which could be conducted from a mobile phone. Clinical decisions were happening on platforms like Microsoft Teams and Zoom,

which allowed patients with mobile phones to connect with clinicians from all over the world. This created an effective opportunity for mPlatform data discovery for network companies allowing multiple ecosystems to emerge and flourish that will hopefully support clinical environments during future epidemics.

Summary and the Future of mHealth

As we continue to evolve into *Homo nodi*, the global adoption of technologies like 5G and subsequent 6G–7G technologies together with improvements in computational speedup and quantum efficiency will see the ability to perform higher fidelity simulations and automations through quantum machine learning and artificial intelligence on small form-factor devices like mobile phones and tablets [92]. Imagine exponential data transfer speeds several orders of magnitude greater than 20gbs speeds that will allow quantum computational renderings of molecular structures for various biochemical and pathological processes to be diagnosed at our fingertips. Not only will it be possible to zoom into molecular renderings when studying complex genetic disorders, but federated learning methodologies will also be considerably easier. Cross-collaboration and platform data analytics will allow for seamless integration between medical organizations, government, and health ministries. Data governance and generalized data protection will be the only limitation for cross-border collaboration needing consortia to manage information flow. However, we are continually seeing the blurring between borders for medical informatics with systems like EPIC and CERNER systems allowing for mobile app integration and accessibility, leading the charge and ensuring seamless data access and electronic health record management across organizations.

Have we reached *Homo Digitalis* status and will we see a blending of human with artificial intelligences? Well it no longer seems a distant future, given the fact that devices are now available that can digitally monitor internal processes and manage diseases. An example of this is the US

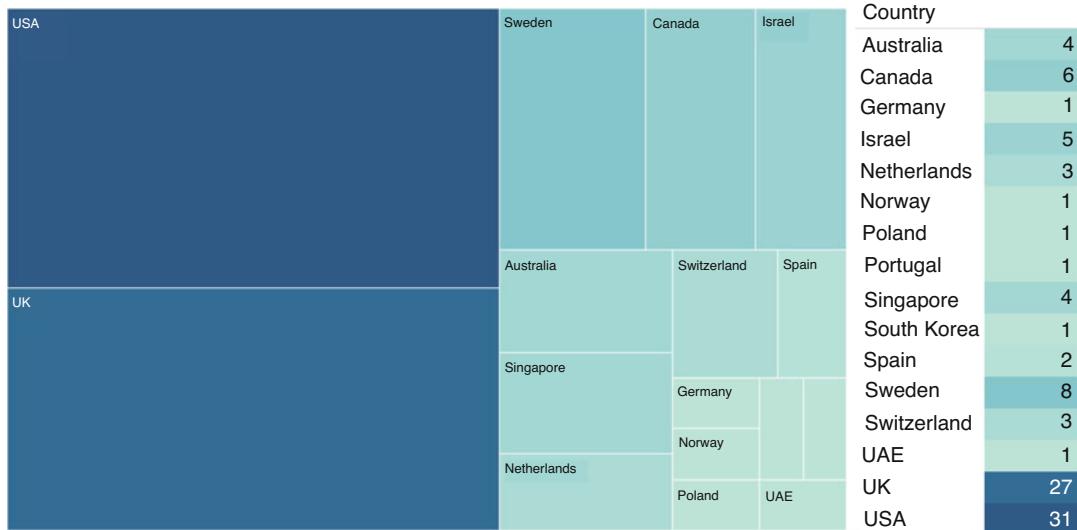


Fig. 4 A sampled distribution of mHealth and AI companies, pooled data from the 2020 Global Digital Health 100 summit illustrating relative contributions within this

FDA-approved Eversense smart fluorescent sub-dermally implantable sensor (made by Senseonics Inc.) that is capable of continuously monitoring blood glucose levels in diabetic patients and forwarding this information to companion smartphone and wearable devices, which can in turn alert the patient if their blood glucose level is too high or too low [93]. mHealth and AI may in the future also facilitate our eventual evolution into *Homo Deus* paving the way to the optimal health we seek [94].

For completeness Fig. 4 summarizes the global distribution of 98 companies making significant progress within the mHealth, eHealth, smartphone, and apps for medicine space. The future for mHealth and AI in medicine is very bright.

space in the developed world. A tableau dashboard presents a non-exhaustive list of companies in the mHealth and AI space based in the UK and abroad [95]

3. Bhalla M, Bhalla A. Generations of mobile wireless technology: a survey. *Int J Comput Appl.* 2010;5(4):26–31.
4. Meraj ud in Mir MaK, S. Evolution of mobile wireless technology from 0G to 5G. *Int J Comput Sci Inf Technol.* 2015;6(3):2545–51.
5. Bainbridge Newsletter. From 1G to 5G: a brief history of the evolution of mobile standards. 2020. <https://www.bainbridge.be/news/from-1g-to-5g-a-brief-history-of-the-evolution-of-mobile-standards>
6. Singh S. Leading trends in information technology [Internet]: Stanford MS&E 238 Blog. 2017. [cited 2021]. <https://mse238blog.stanford.edu/2017/07/ssound/1g-2g-5g-the-evolution-of-the-gs/>
7. Tapper J. London hospital starts virtual ward rounds for medical students. The Guardian; 2020. <https://www.theguardian.com/society/2020/jul/04/london-hospital-starts-virtual-ward-rounds-for-medical-students>
8. Vindrola C, Fulop N, Greenhalgh T. Virtual wards: caring for COVID-19 patients at home could save lives. 2020 November 6, 2020.
9. Museum S. Titanic, Marconi and the wireless telegraph [Webpage]. 2018 [updated 24 October 2018]. <https://www.sciencemuseum.org.uk/objects-and-stories/titanic-marconi-and-wireless-telegraph>
10. Foote K. Data topics [Internet]: Dataversity. 2016. [cited 2021]. <https://www.dataversity.net/brief-history-internet-things/#>
11. Relations CoF. The origins of the Internet. 2017. <https://world101.cfr.org/global-era-issues/cyberspace-and-cybersecurity/origins-internet>
12. Science M, Kuniaevsky M, Blackberry, Apple, Fernandez R, Microsoft, et al. Figure 1 Weblinks 2020. <https://www.sciencemuseum.org.uk/objects-and-stories/>

References

1. Aydin G, Silahtaroglu G. Insights into mobile health application market via a content analysis of marketplace data with machine learning. *PLoS One.* 2021;16(1):e0244302.
2. Kooman JP, Wieringa FP, Han M, Chaudhuri S, van der Sande FM, Usvyat LA, et al. Wearable health devices and personal area networks: can they improve outcomes in haemodialysis patients? *Nephrol Dial Transplant.* 2020;35(Suppl 2):ii43–50.

- titanic-marconi-and-wireless-telegraph. <https://www.sciencemuseum.org.uk/objects-and-stories/invention-mobile-phones>. <https://crackberry.com/evolution-blackberry-pictures>. <https://www.samsung.com/uk/smartphones/galaxy-s21-5g>. <https://www.apple.com/uk/watch/>. <https://www.wired.com/review/focals-by-north-smart-glasses/>. <https://www.microsoft.com/en-us/hololens>
13. O'Dea S, Johnson J. Number of mobile subscriptions worldwide 1993–2019 Statistica; 2019/2020. <https://www.statista.com/aboutus/our-research-commitment>. Statistica Dec 3, 2020. <https://www.statista.com/statistics/249562/number-of-worldwide-internet-users-by-region/>. 31 Jan 2020.
 14. Internet Subscriberships and Users; Mobile cellular subscriptions, Fixed telephone subscriptions, [Internet]. 2020. https://data.worldbank.org/indicator/IT.NET.SECR.P6?end=2019&name_desc=false&start=1980&view=chart
 15. WHO. mHealth New horizons for health through mobile technologies. Based on the findings of the second global survey on eHealth. Global observatory for eHealth series, vol. 3. WHO; 2011. p. 1–104.
 16. Kalan R, Wiysonge C, Ramafuthole T, et al. Mobile phone text messaging for improving the uptake of vaccinations: a systematic review protocol. *BMJ Open*. 2014;4:e005130. <https://doi.org/10.1136/bmjopen-2014-005130>.
 17. BMJ. How can mobile technology improve health in low- and middle-income countries. 2020. <https://beta-www.bmjjournals.com/company/who-we-are/case-studies/mhealth-case-study/>
 18. Clark J, Glasziou P, Del Mar C, Bannach-Brown A, Stehlik P, Scott A. A full systematic review was completed in 2 weeks using automation tools: a case study. *J Clin Epidemiol*. 2020;21:81–90.
 19. Clark J, Sanders S, Carter M, Honeyman D, Cleo G, Auld Y, et al. Improving the translation of search strategies using the Polyglot Search Translator: a randomized controlled trial. *J Med Libr Assoc*. 2020;108(2):195–207.
 20. Mourad Ouzzani HH, Fedorowicz Z, Elmagarmid A. Rayyan – a web and mobile app for systematic reviews. *Syst Rev*. 2016;5:210.
 21. Woo M. An AI boost for clinical trials. *Nature*. 2019;573:S100.
 22. Free C, et al. The effectiveness of mobile-health technologies to improve health care service delivery processes: a systematic review and meta-analysis. *PLoS Med*. 2013;10:e1001363. <https://doi.org/10.1371/journal.pmed.1001363>.
 23. TCGA. The Cancer Genome Atlas for Genomics 2020. <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga>
 24. Leung M, Delong A, Alipanahi B, et al. Machine learning in genomic medicine: a review of computational problems and data sets. *Proc IEEE*. 2016;104(1):176–97.
 25. mymHealth. A real-world service evaluation of myHeart: evaluation MMH-E01: mymHealth. <https://mymhealth.com/studies/real-world-evaluation-myheart-mmh-e01>
 26. Carrasco-Hernandez L, Jódar-Sánchez F, Núñez-Benjumea F, Moreno Conde J, Mesa González M, Civit-Balcells A, et al. A mobile health solution complementing psychopharmacology-supported smoking cessation: randomized controlled trial. *JMIR Mhealth Uhealth*. 2020;8(4):e17530.
 27. Hors-Fraile S, Malwade S, Spachos D, Fernandez-Luque L, Su CT, Jeng WL, et al. A recommender system to quit smoking with mobile motivational messages: study protocol for a randomized controlled trial. *Trials*. 2018;19(1):618.
 28. Norouzi S, Kamel Ghalibaf A, Sistani S, Banazadeh V, Keykhaei F, Zareishargh P, et al. A mobile application for managing diabetic patients' nutrition: a food recommender system. *Arch Iran Med*. 2018;21(10):466–72.
 29. Van Hamme T, Garofalo G, Argones Rúa E, Preuveneers D, Joosen W. A systematic comparison of age and gender prediction on IMU sensor-based gait traces. *Sensors (Basel)*. 2019;19(13):2945.
 30. Mohammadi R, Atif M, Centi AJ, Agboola S, Jethwani K, Kvedar J, et al. Neural network-based algorithm for adjusting activity targets to sustain exercise engagement among people using activity trackers: retrospective observation and algorithm development study. *JMIR Mhealth Uhealth*. 2020;8(9):e18142.
 31. Chaix B, Bibault JE, Pienkowski A, Delamon G, Guillemaillé A, Nectoux P, et al. When chatbots meet patients: one-year prospective study of conversations between patients with breast cancer and a chatbot. *JMIR Cancer*. 2019;5(1):e12856.
 32. Chan S, Li L, Torous J, Gratzer D, Yellowlees PM. Review and implementation of self-help and automated tools in mental health care. *Psychiatr Clin North Am*. 2019;42(4):597–609.
 33. Chong NK, Chu Shan Elaine C, de Korne DF. Creating a learning televillage and automated digital child health ecosystem. *Pediatr Clin North Am*. 2020;67(4):707–24.
 34. Denecke K, Warren J. How to evaluate health applications with conversational user interface? *Stud Health Technol Inform*. 2020;270:976–80.
 35. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR Mhealth Uhealth*. 2018;6(11):e12106.
 36. Issom DZ, Rochat J, Hartvigsen G, Lovis C. Preliminary evaluation of a mHealth coaching conversational artificial intelligence for the self-care management of people with sickle-cell disease. *Stud Health Technol Inform*. 2020;270:1361–2.
 37. Kelly JT, Collins PF, McCamley J, Ball L, Roberts S, Campbell KL. Digital disruption of dietetics: are we ready? *J Hum Nutr Diet*. 2021;34(1):134–46.
 38. Kretzschmar K, Tyroll H, Pavarini G, Manzini A, Singh I. Can your phone be your therapist? Young people's ethical perspectives on the use of fully automated conversational agents (Chatbots) in mental

- health support. *Biomed Inform Insights.* 2019;11:1178222619829083.
39. Loveys K, Fricchione G, Kolappa K, Sagar M, Broadbent E. Reducing patient loneliness with artificial agents: design insights from evolutionary neuropsychiatry. *J Med Internet Res.* 2019;21(7):e13664.
40. Marcus JL, Sewell WC, Balzer LB, Krakower DS. Artificial intelligence and machine learning for HIV prevention: emerging approaches to ending the epidemic. *Curr HIV/AIDS Rep.* 2020;17(3):171–9.
41. Müschenich M, Wamprecht L. [Health 4.0 – how are we doing tomorrow?]. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz.* 2018;61(3):334–9.
42. Nadarzynski T, Bayley J, Llewellyn C, Kidsley S, Graham CA. Acceptability of artificial intelligence (AI)-enabled chatbots, video consultations and live webchats as online platforms for sexual health advice. *BMJ Sex Reprod Health.* 2020;46(3):210–7.
43. Pereira J, Díaz Ó. Using health chatbots for behavior change: a mapping study. *J Med Syst.* 2019;43(5):135.
44. Powell J. Trust me, I'm a chatbot: how artificial intelligence in health care fails the Turing test. *J Med Internet Res.* 2019;21(10):e16222.
45. Tielman ML, Neerincx MA, Pagliari C, Rizzo A, Brinkman WP. Considering patient safety in autonomous e-mental health systems – detecting risk situations and referring patients back to human care. *BMC Med Inform Decis Mak.* 2019;19(1):47.
46. Tudor Car L, Dhingaran DA, Kyaw BM, Kowatsch T, Joty S, Theng YL, et al. Conversational agents in health care: scoping review and conceptual analysis. *J Med Internet Res.* 2020;22(8):e17158.
47. Zhang J, Oh YJ, Lange P, Yu Z, Fukuoka Y. Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet: viewpoint. *J Med Internet Res.* 2020;22(9):e22845.
48. Wasserlauf J, You C, Patel R, et al. Smartwatch performance for the detection and quantification of atrial fibrillation. *Circ Arrhythm Electrophysiol.* 2019;12:e006834.
49. de Marvao A, Dawes T, Howard J, et al. Artificial intelligence and the cardiologist: what you need to know for 2020. *Heart.* 2020;106:399–400.
50. Crooks M, et al. Evidence generation for the clinical impact of myCOPD in patients with mild, moderate and newly diagnosed COPD: a randomised controlled trial. *ERJ Open Res.* 2020;6:00460–2020.
51. van der Heijden M, Lucas PJ, Lijnse B, Heijdra YF, Schermer TR. An autonomous mobile system for the management of COPD. *J Biomed Inform.* 2013;46(3):458–69.
52. North M, Bourne S, Green B, et al. A randomised controlled feasibility trial of E-health application supported care vs usual care after exacerbation of COPD: the RESCUE trial. *npj Digit Med.* 2020;3:145.
53. Rajpurkar P, et al. CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv:171105225.* 2017.
54. Tiron R, Lyon G, Kilroy H, Osman A, Kelly N, O'Mahony N, et al. Screening for obstructive sleep apnea with novel hybrid acoustic smartphone app technology. *J Thorac Dis.* 2020;12(8):4476–95.
55. HBP. The human Brain Project. <https://www.humanbrainproject.eu/en/science/overview/>
56. NIH. Human Connectome Project: National Institute of Health. <http://www.humanconnectomeproject.org/about/links/>. <https://neuroscienceblueprint.nih.gov/human-connectome/connectome-projects>
57. Barbour K, Hesdorffer D, Tian N, et al. Automated detection of sudden unexpected death in epilepsy risk factors in electronic medical records using natural language processing. *Epilepsia.* 2019;60:1209–20.
58. Berrouiguet S, Barrigón ML, Castroman JL, et al. Combining mobile-health (mHealth) and artificial intelligence (AI) methods to avoid suicide attempts: the Smartcrises study protocol. *BMC Psychiatry.* 2019;19:277.
59. Hügle M, Omoumi P, van Laar J, Boedecker J, Hügle T. Applied machine learning and artificial intelligence in rheumatology. *Rheumatol Adv Pract.* 2020;4(1):rkaa005. Published 2020 Feb 19. <https://doi.org/10.1093/rap/rkaa005>.
60. Kothari S, et al. Artificial Intelligence (AI) and rheumatology: a potential partnership. *Rheumatology.* 2019;58(11):1894–5.
61. Symptomate. Symptomate Symptom Checker. <https://symptomate.com/chatbot/>
62. Isabel. Isabel Symptom checker website. <https://symptomchecker.isabelhealthcare.com/>
63. Shiezaedeh Z, Sajedi H, Afkakie E. Diagnosis of rheumatoid arthritis using an ensemble learning approach. In: Computer science & information technology (CS & IT). San Jose: Academy & Industry Research Collaboration Center (AIRCC), Horizon Research Publishing; 2015. p. 139–48.
64. Zhou S, UK Biobank Follow-up and Outcomes Group, Brophy S, et al. Defining disease phenotypes in primary care electronic health records by a machine learning approach: a case study in identifying rheumatoid arthritis. *PLoS One.* 2016;11(5):e0154515.
65. Chin C, Hsieh S, Tseng V. eDRAM: effective early disease risk assessment with matrix factorization on a large-scale medical database: a case study on rheumatoid arthritis. *PLoS One.* 2018;13(11):e0207579.
66. Vodenarevic A, van der Goes M, Medina O, et al. Predicting flare probability in rheumatoid arthritis using machine learning methods. In: Proceedings of the 7th international conference on data science, technology and applications, Hampshire, UK: SCITEPRESS. Science and Technology Publications; 2018. p. 187–92.
67. Ceccarelli F, Sciandrone M, Perricone C, Galvan G, Cipriano E, Galligari A, et al. Biomarkers of erosive arthritis in systemic lupus erythematosus: application of machine learning models. *PLoS One.* 2018;13(12):e0207926.

68. Le Berre C, et al. Application of artificial intelligence to gastroenterology and hepatology. *Gastroenterology*. 2020;158(1):76–94.e2.
69. Săftoiu A, Vilmann P, Gorunescu F, et al. Efficacy of an artificial neural network-based approach to endoscopic ultrasound elastography in diagnosis of focal pancreatic masses. *Clin Gastroenterol Hepatol*. 2012;10:84–90.e1.
70. Chen J, Remulla D, Nguyen J, et al. Current status of artificial intelligence applications in urology and their potential to influence clinical practice. *BJU Int*. 2019 Jun 20. Epub ahead of print. Erratum in: *BJU Int*. 2020 Nov;126(5):647. PMID: 31219658. <https://doi.org/10.1111/bju.14852>.
71. Davids J, Savvas-George M, Ashrafiyan H, Darzi A, Marcus H, Giannarou S. Automated vision-based microsurgical skill analysis in neurosurgery using deep learning: development and preclinical validation. *World Neurosurg*. 2021;149:e669. Accepted for Publication.
72. Tapiero S, Yoon R, Jefferson F, et al. Smartphone technology and its applications in urology: a review of the literature. *World J Urol*. 2020;38:2393–410.
73. Danaei G, et al. National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2·7 million participants. *Lancet*. 2011;378(9785):31–40.
74. Shenoy VN, Aalami OO. Utilizing smartphone-based machine learning in medical monitor data collection: seven segment digit recognition. *AMIA Annu Symp Proc*. 2017;2017:1564–70.
75. mymHealth. A real-world multi-centre service evaluation of myDiabetes mymHealth. 2020. <https://mymhealth.com/studies/real-world-multi-centre-evaluation-mydiabetes-mmh-e02>
76. Savala R, Dey P, Gupta N. Artificial neural network model to distinguish follicular adenoma from follicular carcinoma on fine needle aspiration of thyroid. *Diagn Cytopathol*. 2018. <https://doi.org/10.1002/dc23880>.
77. Esteva A, Kuprel B, Novoa R, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115–8.
78. Brinker TJ, Hekler A, Enk AH, Berking C, Haferkamp S, Hauschild A, et al. Deep neural networks are superior to dermatologists in melanoma image classification. *Eur J Cancer*. 2019;19:11–7.
79. Emin E, Emin E, Papalois A, Willmott F, Clarke S, Sideris M. Artificial intelligence in obstetrics and gynaecology: is this the way forward? *In Vivo*. 2019;33(5):1547–51.
80. Tsur A, et al. Development and validation of a machine learning model for prediction of shoulder dystocia. *Ultrasound Obstet Gynecol*. 2019;56. <https://doi.org/10.1002/uog.21878>.
81. Ayres-de-Campos D, et al. SisPorto 2.0: a program for automated analysis of cardiotocograms. *J Maternal-Fetal Med*. 2000;9:311–8.
82. Haijing T, Wang T, Li M, Yang X. The design and implementation of cardiotocography signals classification algorithm based on neural network. *Comput Math Methods Med*. 2018;2018:Article ID 8568617, 12 pages.
83. Porter P, Abeyratne U, Swarnkar V, Tan J, Ng TW, Brisbane JM, et al. A prospective multicentre study testing the diagnostic accuracy of an automated cough sound centred analytic system for the identification of common respiratory disorders in children. *Respir Res*. 2019;20(1):81.
84. Kazi AM, Qazi SA, Khawaja S, Ahsan N, Ahmed RM, Sameen F, et al. An artificial intelligence-based, personalized smartphone app to improve childhood immunization coverage and timelines among children in Pakistan: protocol for a randomized controlled trial. *JMIR Res Protoc*. 2020;9(12):e22996.
85. Chan K, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. *JMIR Med Educ*. 2019;5(1):e13930.
86. Dias R, Gupta A, Yule S. Using machine learning to assess physician competence: a systematic review. *Acad Med*. 2019;94(3):427–39.
87. Setyonugroho W, Kennedy K, Kropmans T. Reliability and validity of OSCE checklists used to assess the communication skills of undergraduate medical students: a systematic review. *Patient Educ Couns*. 2015;98(12):1482–91.
88. Ryan P, Luz S, Albert P, Vogel C, Normand C, Elwyn G. Using artificial intelligence to assess clinicians' communication skills. *BMJ*. 2019;364:l161.
89. Butow P, Hoque E. Using artificial intelligence to analyse and teach communication in healthcare. *Breast*. 2020;50:49.
90. Jani K, Jones K, Jones G, Amiel J, Barron B, Elhadad N. Machine learning to extract communication and history-taking skills in OSCE transcripts. *Med Educ*. 2020;54:1159–70.
91. Spencer G. Defeating dengue fever: AI boosts the global fight against mosquito-borne diseases: Microsoft; 2020. <https://news.microsoft.com/apac/features/defeating-dengue-fever-ai-boasts-the-global-fight-against-mosquito-borne-diseases/>
92. Masters K. Artificial intelligence in medical education. *Med Teach*. 2019;41(9):976–80.
93. FDA. Eversense continuous glucose monitoring system – P160048/S006 Food and Drugs Administration: Food and Drugs Administration. <https://www.fda.gov/medical-devices/recently-approved-devices/eversense-continuous-glucose-monitoring-system-p160048s006>
94. Harari YN. *Homo Deus: a brief history of tomorrow*. London: Vintage; 2017.
95. Global Health. Global digital health 100 announcing the 2020 global digital health 100; Recognising innovation and emerging technologies in healthcare. 2020. <https://thejournalofmhealth.com/digital-health-100/>



Zixin Shu, Ting Jia, Haoyu Tian, Dengying Yan, Yuxia Yang, and Xuezhong Zhou

Contents

Introduction	1248
Methodological Approaches	1249
Data Sources	1249
Knowledge Engineering Approaches	1250
Machine Learning and Data Mining Approaches	1251
Applications and Related Work	1252
Shortcomings and Future Directions	1256
Conclusion	1257
References	1258

Abstract

Alternative medicine (AM) is one of the medical fields that use more natural and traditional therapies for disease diagnosis and treatment, in which traditional Chinese medicine (TCM) now has been recognized as one of the main approaches of AM. As a clinical and evidence-driven discipline with long histories, AM is also heavily relied on in the utilization of big healthcare and therapeutic data for improving

the capability of diagnosis and treatment. In particular, artificial intelligence (AI) has been widely adopted in AM to deliver more practical and feasible intelligent solutions for clinical operations since 1970s. This chapter summarizes the main approaches, related typical applications, and future directions of AI in AM to give related researchers a brief useful reference. We find that although AM has not been widely used in clinical practice internationally, the AI studies showed abundant experiences and technique trials in expert system, machine learning, data mining, knowledge graph, and deep learning. In addition, various types of data, such as bibliographic literatures, electronic medical records, and images were used in the related AI tasks and studies. Furthermore, during this COVID-19 pandemic era, we have witnessed the clinical effectiveness of TCM for COVID-19 treatment, which

Zixin Shu, Ting Jia, Haoyu Tian, Dengying Yan and Yuxia Yang contributed equally with all other contributors.

Z. Shu · T. Jia · H. Tian · D. Yan · Y. Yang · X. Zhou (✉)
Institute of Medical Intelligence, School of Computer and
Information Technology, Beijing Jiaotong University,
Beijing, China
e-mail: xzzhou@bjtu.edu.cn

mostly was detected by real-world data mining applications. This indicates the potential opportunity of the booming of AI research and applications in various aspects (e.g., effective clinical therapy discovery and network pharmacology of AM drugs) in AM fields.

Keywords

Artificial intelligence · Machine learning · Knowledge engineering · Alternative medicine · Traditional Chinese medicine · Clinical therapy · Clinical diagnosis · Network pharmacology · Syndrome · Complementary and alternative medicine

Introduction

Alternative medicine (AM) or complementary and alternative medicine (CAM) is the field for medical products and practices that are not part of mainstream biomedical care [1]. Actually, before the establishment of modern biomedical science in the eighteenth century, the current therapies (e.g., herbal prescriptions) used in AM were the mainstream approaches for healthcare management. People now may use the term “natural,” “holistic,” “home remedy,” or “Eastern Medicine” to refer to AM. However, experts often use five regular related categories to describe it, which are mainly related to treatment or therapies: mind-body therapies (e.g., meditation, hypnosis, and yoga), biologically based practices (e.g., vitamins and dietary supplements, herbs, and special food), manipulative and body-based practices (e.g., massage, chiropractic therapy), biofield therapy (e.g., reiki and therapeutic touch), and whole medical systems (e.g., Ayurvedic medicine, traditional Chinese medicine (TCM), and homeopathy) [2, 3]. AM is used along with standard medical treatment but is not considered by itself to be standard treatment, for example using acupuncture to help lessen some side effects of cancer treatment or a special diet to treat cancer instead of cancer drugs that are prescribed by oncologists [4, 5].

From 2014, the National Cancer Institute (NCI) and the National Centre for Complementary and

Integrative Health (NCCIH) have been sponsoring or cosponsoring clinical trials that test AM treatments and therapies in people, which include 65 related cancer AM trials (<https://www.cancer.gov/about-cancer/treatment/clinical-trials/cam-procedures>). Evidence suggests that several herbal medicines and dietary supplements can alleviate the side effects of cancer treatments, such as nausea, pain, and fatigue [6]. In addition, TCM has been increasingly adopted as a complementary medical therapy for various kinds of diseases, such as cancer [7], rheumatoid arthritis [8, 9], migraine, and functional disorders [10]. As of 2015, there were 3966 TCM hospitals and 45,528 TCM clinics across China (<https://www.who.int/westernpacific/health-topics/traditional-complementary-and-integrative-medicine>). In this COVID-19 pandemic, as a typical effective solution for virus-related infections, TCM plays a significant role in the relative low mortality rate of COVID-19 inpatients amid the situation of heavy healthcare overload during early days in Feb–Apr 2020 [11]. For example, a recent retrospective cohort study had shown the clinical effectiveness of add-on herbal prescriptions for COVID-19 patients, which even largely reduced the mortality of COVID-19 severe inpatient cases [12, 13].

Due to the empirical nature of medical science, the discovery and updating of clinical evidence and solutions are heavily relied on in the utilization of various kinds of clinical data. Therefore, biomedical science needs artificial intelligence (AI) essentially to assist the generation of new knowledge and decision-making for clinical operations. Since the early work of medical AI (e.g., MYCIN [14]) in the 1970s, the related AI methodologies in medicine have evolved from production rules to data mining and now to deep learning [15, 16], including AI techniques such as fuzzy expert systems, Bayesian networks, artificial neural networks, and hybrid intelligent systems used in different clinical settings in healthcare. Since the recent ten years, the most popular medical AI techniques have been machine learning and deep learning for clinical diagnosis [17], treatment recommendation [18], and healthcare management [19], which recruit big chunks of investments, owing to their higher social and business values compared with other

fields. The related historical events, trends, and techniques are similar for AM. With the first bunch of work on medical expert systems since the 1970s, now the main AI approaches adopted in AM are data mining and deep learning, with particular applications on decision support system and precision medicine [20, 21].

Methodological Approaches

Data Sources

AI applications in AM are heavily relied on in available data sources. Typically, several types of data sources, such as bibliographic literature, curated structured databases, terminological systems, and clinical databases have significant contributions to the development of AI applications and methods in AM fields. Here we will introduce the information fields of the data sources by taking classic database examples.

First, bibliographic literature, which is published by medical researchers, holds significant knowledge for various medical fields, including AM as one of the subsets. PubMed proposes the largest data collection of bibliographic literature, which is a high valuable data source for AI applications. The bibliographic literature records include the fields title, abstract, full text (partially), Medical Subject Headings (MeSH), and other biometric fields (e.g., authors, publication date, and citations), which deliver a large-scale data source for medical text mining and the curation of subject-specific structured databases (e.g., drug-indication association database).

In the TCM field, a similar bibliographic literature database, the Traditional Chinese Medical Literature Analysis and Retrieval System (TCMLARS), contains more than one million records of traditional Chinese medicine literature since 1949. It is a large-scale traditional Chinese medicine database established by the Institute of Information on TCM of the China Academy of Chinese Medical Sciences, with a corresponding English version [22, 23].

Second, many subject-specific structured databases, such as herbal database, prescription

database, disease database, and pharmacological databases, were curated from textbooks, AM dictionaries, and bibliographic literature database. The content of these databases consists of both structured and text fields. For example, herbal database would contain typical fields such as herbal name, synonyms, indication, herb-related ingredients, and related studies. In recent years, pharmacological databases, such as TCMSP [24], ETCM [25], HERB [26], and SymMap [27], which have herb-related pharmacological information, were curated and proposed as a significant data source for network pharmacological research. For example, SymMap contains 499 herbal medicines and 1717 TCM symptoms derived from the 2015 edition of Chinese Pharmacopoeia and maps TCM symptoms to 961 Western symptoms. At the same time, 5235 diseases associated with these symptoms; 19,595 herbal ingredients; 4302 drug targets; and the correlation between these six types of data were included. SymMap links traditional Chinese medicine and modern medicine from a phenotype to a molecular level, which has been increasingly used for network pharmacological studies.

Third are terminological systems, including controlled vocabularies and medical ontologies, that propose the semantic mappings of medical terms and entities in different data sources. For example, in 2000s, a TCM medical ontology named the unified traditional Chinese medical language system (UTCMLS) was developed, which supports TCM language knowledge storage, concept-based information retrieval, and information integration for bibliographic records [28]. In order to solve the problem of expression difference in TCM terms, a clinical terminology system similar to SNOMED CT was developed to standardize the terms and expressions of clinical TCM concepts, which would be an important terminology database for AI research in TCM [29]. With the formal introduction of the concept of knowledge graph by Google in 2012, the knowledge graph for alternative medicine has been gradually developed. For example, the traditional Chinese medicine language system (TCMLS) semantic network, including 127 semantic types and 58 semantic relations, and knowledge graph of TCM characteristic therapy, including

auricular point therapy, moxibustion therapy, medicinal wine therapy, cupping therapy, and bloodletting therapy, were developed in recent years.

Fourth, clinical case database is one of the most significant data sources for medical knowledge discovery and clinical decision support, which would often not be publicly available, owing to ethnic and privacy issues. For example, in 2010, Zhou et al. [21] developed a clinical data warehouse that incorporates the structured electronic medical record (SEMR) data for medical knowledge discovery and TCM clinical decision support (CDS) and integrates 20,000 TCM inpatient data and 20,000 outpatient data.

Knowledge Engineering Approaches

Knowledge engineering (KE) is one of the substantial techniques of AI [30], which has been developed during the early days of AI's timeline. This type of AI technique also has been adopted as the main approach for AI in medicine to develop a knowledge base and expert system for clinical decision support. It is even important for AI in AM since the AM field involves practical issues related to various terminological representations of medical concepts and rich empirical knowledge from even ancient literature for disease treatment. In the following, we would introduce the main KE techniques adopted in the field of AI in AM.

Expert System

Since 1970s, in AM field, researchers had tried to collect expert knowledge and build knowledge base to simulate the decision process of AM experts. For example, Guan Youbo's computer program for the diagnosis and treatment of liver diseases, the first TCM expert system in China, has played an important role in the development of TCM diagnosis and the treatment decision support system since it came out in 1979 [31]. After that, hundreds of TCM expert systems with rule reasoning and machine learning as the main methods have been developed. The main function of these systems is to support TCM diagnosis, including some general TCM expert system

development tool software, such as GTS (1985) [32], Monkey (1988) [33], YHW-CTMEST (1988) [34], etc. Similarly, in 2014, Nopparakiat et al. developed a rule-based expert consultation system for skin problems, which combines Thai traditional medical knowledge to propose Thai herbal formula and Thai herbal cosmetics for the treatment of skin problems [35].

Ontology and Knowledge Graph

Medical ontology and knowledge graph have been focused by AI researchers since large amount of empirical knowledge from textbooks and ancient literatures is available in AM. Furthermore, due to the various types of data sources and individualized treatment, the terms on diagnosis and treatment are often nonstandardized and diverse [36]. Knowledge representation methods provide a formal semantic approach for traditional medical knowledge, which would help to preserve knowledge in this field and share common understanding among different clinical communities and groups. Therefore, many regions, for example, China, India, Greece, South Korea, Indonesia, Thailand, and the West Africa, have developed their own medical ontologies. The unified traditional Chinese medicine language system (UTCMLS) [28] is a system developed by 16 Chinese medicine universities and hundreds of researchers, which promotes the development of TCM knowledge base through ontology methodology and structure design. The traditional Chinese medicine language system (TCMLS) [37] is an integrated system of Chinese medicine language based on the subject system of traditional Chinese medicine. It covers the subject system of traditional Chinese medicine and the natural and human science vocabularies related to traditional Chinese medicine, such as biology, plant and chemical engineering, etc. On the other hand, the knowledge graph of TCM characteristic therapy systematically integrated the knowledge of TCM characteristic therapy, including auricular point therapy, moxibustion therapy, wine therapy, cupping therapy, and bloodletting therapy [38].

In a study conducted by Raja Mohan and Arumugam, an ontology was developed and inferred for Indian medicinal plants using the

Protégé software [39]. Hyunchul Jang et al. started with medicinal materials to express the relationship between symptoms, disease, and the treatment of patients and constructed an ontology of traditional Korean medicine (TKM) [40]. SysMEDTRAD is a West African traditional medicine management system based on ontoMEDTRAD (an ontology of West African TM associated with a visual approach), which automatically generates iconic recipes of plant-based medicines from the ontology [41].

Machine Learning and Data Mining Approaches

To discover the hidden knowledge and propose automatic clinical decision capabilities for disease management, various kinds of data mining and machine learning methods were used in AI in the AM field, which has mainly received its attention since the 2000s. For well-defined and structured data, besides the traditional methods (e.g., decision tree, Native Bayes, and regression) that were widely used in AI in medicine, typically data mining methods, such as association rule, frequent itemset, and complex network, and machine learning methods, including classification methods (e.g., support vector machine, Bayesian network, multilabel learning, and deep learning) and clustering methods (e.g., latent class model and primary component analysis) were frequently applied in AM diagnosis and treatment analysis tasks. In addition, due to the significant role of electronic medical record for the storing of clinical manifestations and diagnoses of AM clinical procedures, text mining or information extraction has been a rising focus task for AM clinical data analysis. However, in AM field, time series analysis and image processing were relatively less studied due to the lack of data availability.

At the beginning of the twenty-first century, more AI systems incorporating machine learning methods have emerged. For example, the combination of multiple data mining techniques mainly used improved hybrid Bayesian network learning techniques to build a new type of TCM diagnosis self-learning expert system, which has achieved

encouraging results [42]. In order to fully demonstrate Thailand's traditional medical knowledge, a rule-based expert system was researched and developed for skin problem consultation [35], which helped promote Thai traditional medicine. In addition, machine learning algorithm such as support vector machine (SVM) [43], Multiple Asymmetric Partial Least Squares Classifier (MAPLSC) [44], Naive Bayes [45], K-nearest neighbor (KNN) [46], and other methods were used to diagnose the collected clinical images [47–51], especially in the tongue and vein field [50]. Association rule learning in data mining is a frequently used method. This method has been applied to the field of Chinese medicine since 2000 and is still widely used. The most widely used field is the study of the combination regularities of herbal prescriptions. Yao et al. [52] applied association rule analysis technology to conduct exploratory analysis and research on the scientific connotation of Chinese herbal compound compatibility for the treatment of diabetes. Huang et al. [53] used modern statistics and data mining techniques to analyze and excavate acupuncture and moxibustion treatment of insomnia.

Cluster analysis is the use of mathematical methods to study and process the classification of a given object, which reflects the general rule “Things are gathered by clusters, and people are divided by groups” [54] and proposes a significant machine learning approach for AM data analysis. Cluster analysis in the AM field began to develop at the end of the twentieth century. For example, Qi et al. [55] used the fuzzy cluster analysis to study the differences between different species of *Scutellaria baicalensis* Georgi and discussed the reasonable selection of substitutes and obtained satisfactory results. Furthermore, cluster analysis is also widely used in the detection of syndrome groups, symptom clusters, and drug combinations. For example, Wang et al. [56] applied the variable cluster analysis to detect the TCM syndromes of advanced lung cancer. Tang et al. [57] proposed a complex system entropy clustering method for discovering new Chinese medicine prescriptions. Among the clustering methods, the most important one is the latent tree model, which is a particular type of probabilistic graphical

model developed by Zhang et al. [58]. Wang et al. [59] used the latent tree model to analyze the collected kidney deficiency data sets. The established model is consistent with the theory of TCM: latent variables correspond to syndrome factors, and latent clusters correspond to syndrome types. This indicates that the latent tree model is a feasible and reliable method for detecting the underlying syndrome structures in TCM, which might fit for other AM fields as well. Extracting meaningful information and knowledge from free text is an important research topic in the field of machine learning and data mining, and text mining has become one of the most active research branches in the field of data mining in the AM field [23]. For example, using dictionary-based or rule-based machine learning and statistical methods for named entity recognition (NER), Cao et al. [60] developed an ontology-based system for extracting knowledge on TCM drugs and formulations from semi-structured texts.

The application of deep learning also began to sprout in the AM field in 2015. A deep-learning-based multilabel learning framework was used to effectively process the modeling tasks of TCM data [61]. A Graph Convolutional Network was also applied to obtain the TCM meridian prediction model and discover the relationship between herbal compounds and meridians [62]. In particular, deep representation learning was used for text mining tasks (e.g., NER) often with state-of-the-art performance. The addition of deep learning allows NER to be better developed in electronic medical records. Ji et al. [63] proposed a collaborative cooperation of multiple neural-network-model-based approach, which consists of two BiLSTM-CRF models and a convolutional neural network (CNN) model, obtaining the best performance on NER in Chinese electronic medical records on the 2019 China Conference on Knowledge Graph and Semantic Computing. Furthermore, due to the needs of translational bioinformatics approach to connect the clinical discoveries with their underlying biological molecular mechanisms, a complex network approach was used to integrate real-world clinical data and network medicine data sources for detecting the potential novel clinical disease subtypes. A

complex network approach was used to integrate real-world clinical data and network medicine data sources (e.g., phenotype-genotype associations) for detecting the potential novel clinical disease subtypes. For example, Shu et al. [64] used TCM clinical electronic medical records (EMRs) of 6475 liver inpatients to construct a liver disease comorbid network (LDCN) to investigate the significant clinical subtypes of chronic liver diseases.

Compared with other AM fields, it is obvious that the literature published on AI in TCM, in particular herbal therapies or medicine, has the highest proportion. In other therapies, although there is much less work conducted, data analysis and AI methods were already used for more specific clinical solutions, such as massage [65] and Tai Chi. For example, AI systems were tried to help design a Tai Chi robot control system [66]. The use of AI in AM is also reflected in essential oil therapy. Related work in 2015 proposing to use artificial neural networks to predict the antibacterial activity of essential oils has made essential oil therapy more widely used [67]. In traditional Indian medicine, the “Prakriti” method in the Ayurvedic medical system is verified by using a variety of machine learning methods, proving that the clinical methods of Prakriti evaluation are nonempirical and, further, can be recapitulated and formalised through advanced machine learning approaches. Due to poor applicability and difficulty in collecting data samples, and other various reasons, there is few relevant research in other AM fields.

Applications and Related Work

Clinical Diagnosis

AI has been widely used in the clinical diagnosis of AM for a long time. It includes tongue [68], pulse [69], acupuncture [70, 71], herbal medicine [72], etc. The key aspects of application can be listed as follows: feature extraction [73], clinical decision support [74], and image classification [75]. Many diseases are involved in clinical practice, such as angina pectoris [76], lung cancer [77], polycystic ovary syndrome [78], etc.

With the rapid development of AI, the combination of Chinese medicine and AI will provide

assistance for the inheritance and modernization of Chinese medicine. For example, the colors of the tongue and the thickness of the tongue coating can be identified by a KNN-based feature matching method [75] with two steps (i.e., region separation and color recognition). The experiments showed that the proposed method was simple and highly robust and can be applied to tongue characterization effectively. For the objective research of the four diagnosis of TCM, AI technology can quickly complete the collection of various physical signs of the patient's body and give relevant diagnosis suggestions [74, 79]. Related researchers studied the intelligent model of TCM syndrome diagnosis differentiation through a variety of methods. According to 1134 clinical survey data of patients with type 2 diabetes, Li et al. [80] used artificial neural networks (ANN) and fuzzy systems (FS) to establish an adaptive fuzzy inference system model based on dynamic Kohonen networks. The model was used to mine clinical data, based on the basic theories of TCM, to obtain the diagnostic criteria for common syndromes of type 2 diabetes.

In traditional Indian medicine, Aniruddha Joshi et al. [81] described a system of generating pulse waveforms and used various feature detecting methods to show that an arterial pulse contains typical physiological properties. In Korean medicine, Kwang-Baek Kim et al. [82] proposed a self-diagnosis system of Korean traditional medicine based on Korean Standard Causes of Death Disease Classification Index (KCD) and the fuzzy ART/inference method. It can accept symptoms of a user from a certain part of their body where they feel inconvenient. Then five most probable diseases, with their causes and treatments extracted from Korean traditional medicine books, were picked up based on the fuzzy ART algorithm and fuzzy inference engine. It is verified by field experts in Korean traditional medicine knowledgeable in the collection of symptom-disease-treatment relationships and performance evaluation of experiment results.

Besides the above applications, AI has many other applications in AM. With the development of big data, the Internet, and 5G technology, AI will undoubtedly be more widely used in the

medical and health fields in the future [78], providing more help to humans, and its influence will become more extensive and deeper, and even promote changes in the medical models, reshaping the entire medical industry. However, all sciences and technologies are the crystallization of human wisdom, and medical AI is no exception. While making innovations and breakthroughs, it also continuously produces new problems and new difficulties that people have never encountered before; the related challenges, such as ethical issues, human-computer interaction, subject responsibility, etc., are necessary to review and reflect on a new level of understanding.

Clinical Therapies

The application of AI in clinical therapies is mainly reflected in the area of assisting in the decision-making of clinical treatment. At the end of the 1970s, artificial intelligence was introduced into the field of TCM, which provided advanced productive forces for the modernization of TCM, and the expert system of TCM came into being [83]. However, due to the problem of objectification and standardization of the four-diagnosis information, for a long period, the TCM expert system stayed in the stage of assisting clinical diagnosis and preserving the academic thought of old TCM doctors and did not make substantial progress [84]. Until recent years, the TCM expert system has gradually experienced the development process from TCM prescription analysis to clinical decision support and then to prescription recommendation for specific diseases. In 2007, Liu et al. [85] developed the intelligent analysis system of TCM prescription (CPIAS), which was the first attempt to analyze clinical prescriptions and for the first time introduced the quantitative method through the extraction of four-diagnosis information and prescription from TCM medical cases to carry out a series of prescription analysis, and obtain the pattern of syndrome differentiation, but still did not involve clinical decision-making. In 2011, Zhou et al. [21] developed the TCM clinical data warehouse platform for medical knowledge discovery and decision support based on structured electronic medical record (SEMR)

data. This platform has greatly promoted the development of medical knowledge discovery and TCM clinical decision support. Then in 2015, the Clinical Auxiliary Decision-Making System of Traditional Chinese Medicine TCM-CDS came out [86]. The system used agent technology to process clinical thinking artificially and intelligently and used a large number of machine learning algorithms to realize the system's independent decision and constructed an intelligent expert decision system. Subsequently, the TCM prescription intelligent decision-making system for various diseases (e.g., lung cancer [87], hypertension [88]) began to develop gradually, which greatly improved the clinical decision-making capability of TCM.

Traditional Ayurveda medicine has also developed AI-related systems. As early as 1989, the Resource Center for Indian Language Technology Solutions-Malayalam designed and developed an expert system named Prakes, which provided diet advice, daily health advice, and the possibility of disease and preventive measures [89]. In 2013, the Ayurvedic Medical system released a groundbreaking product, AyuSoft [90], an interactive software that provided end-to-end medical solutions based on Ayurveda and helped give health advice. It has a wide range of applications, including disease diagnosis and treatment, diet and lifestyle advice, personal management information systems, and text and analysis reporting tools. Nowadays, the intelligent system research of Ayurvedic medicine is still in development. The establishment of a theoretical model based on clinical treatment practice is still the future development direction of artificial intelligence.

At present, the application of artificial intelligence in the field of acupuncture research is still in its infancy [91]. In 2007, Chi Fai David Lam developed the Chinese acupuncture expert system (CAES) [92], which can provide a list of relevant diagnoses according to the symptoms and signs of patients and corresponding TCM acupuncture treatment suggestions for users. A software and hardware integration system of TCM acupuncture and moxibustion that appeared in 2012 was mainly used to assist doctors in the operation of clinical acupuncture and moxibustion, filling the

clinical needs for acupuncture and moxibustion in China at that time [93].

An intelligent meridian-assisted diagnosis and treatment system named Academician Cheng Shennong was developed in August 2017, which largely promotes AI systems for acupuncture in the trend of AI [94]. In the same year, Acubots (Digital Meridian Intelligent Acupuncture Robot System [95]), an interuniversity scientific innovation team led by researchers from the Nanjing University of Traditional Chinese Medicine in China, was unveiled. In 2019, a robot-controlled acupuncture (RCA) [96] device made the vision of robot-controlled acupuncture now a reality [97]. However, it is still a long way to deliver a practical AI system for the clinical use of acupuncture therapies.

Physical and mental therapies (e.g., meditation and yoga) have been proved to be an effective AM solution for treating a variety of diseases and improving physical and mental quality. In recent years, researchers usually combine mobile applications with psychosomatic therapy to achieve convenient and practical purposes. Collaborative research in the field of psychosomatic therapy and artificial intelligence in the future is still an opportunity and a challenge for scientific research [98]. In 2015, Vijayaragavan et al. developed an app to perform yoga and music therapy on a person by monitoring his electroencephalogram (EEG) readings; as a result, people are now able to relax and restore their peace of mind under stressful conditions [99]. In 2016, researchers developed a neural adaptive virtual reality meditation system named RelaWorld [100] to combine virtual reality with neural feedback. The system used a head-mounted display, which allowed users to float in the virtual world through meditation exercises and measure the user's brain activity in real-time through EEG. The system provided an easy-to-use tool for beginners and even provides added value for experienced meditators. In 2017, Prasanna M [101] introduced an Android app that has been connected to an Internet-of-Things-based (IoT) hardware system that monitored yoga people's blood pressure, heart-beat, and temperature; helped people perform various yoga postures; and helped sense body

responses. This Android-app-based IoT suite also provided yoga guidance to users and was easy to operate and use.

Pharmacological Applications

The typical example of the integration of artificial intelligence and clinical pharmacology is network pharmacology, which is a new cross-discipline based on multidisciplinary methodologies such as system biology, computer science, and multi-pharmacology. Network pharmacology is devoted to elucidating the action mechanism of drugs from the point of view of a biological network. It is an effective approach to discovering the active substances of TCM and revealing the pharmacological mechanism of TCM [102]. In 2007, Hopkins [103] first put forward the concept of network pharmacology, which broke the traditional concept of “one drug, one target, one disease.” Under the inspiration of system biology, it realized that drugs play a role in the treatment of complex diseases through multiple targets and pointed out that research on biological networks rather than single targets will be the direction of the new generation of drug research. In the same year, Li [104] established a biological network framework for TCM research, which was an important basis for the combination of network pharmacology and TCM. Since then, network pharmacology has rapidly become the forefront of drug research [105]. In 2011, to further solve the major problem of the unknown mechanism of TCM, Li put forward the novel concept of “network target” instead of the traditional “single target” paradigm, which became the main framework for TCM network pharmacology [106].

At present, network pharmacology is often used to help discover the underlying mechanisms of drug action and drug repurposing and the biological mechanisms of disease phenotypes in the TCM field, with various kinds of applications for many different herbal prescriptions or herbal ingredients. Here, we take a few examples for demonstration. Yu et al. [107] obtained the chemical composition and action target of TCM compound prescription in the Yinhuang Qingfei capsule through network pharmacological analysis and found eight main targets for the treatment

of chronic bronchitis. Then based on the previous drug-ingredient-target data, the active components acting on chronic bronchitis in the Yinhuang Qingfei capsule were deduced, and the known drugs with a similar structure to the Yinhuang Qingfei capsule were collected for molecular docking test, and then the active components of the drugs were verified. Taking the Liuwei Dihuang Pill as an example, Wang et al. [108] developed a new network pharmacology method to identify potential new indications of classical compound prescription by integrating drug target protein, chemical structure, action mechanism, and disease-related molecular network. It was found that the Liuwei Dihuang Pill also played a therapeutic role in immune system diseases and tumors, which further expanded the scope of application of traditional prescriptions. Ma et al. [109] integrated literature mining methods to construct a molecular network between hypertensive phenotypic genes and symptomatic phenotypic genes of hyperactivity of liver-yang syndrome. They further carried out a cluster analysis to interpret the biological mechanisms of hyperactivity of liver-yang syndrome in hypertensive disease from the genetic level. Network pharmacology has become an important tool for the development of multitarget new drugs, in particular for such diseases without cure drugs. For example, using integrative network pharmacological methods, Fang et al. [110] excavated ten potential traditional Chinese medicines, such as *Ginkgo biloba*, saw-tooth grass, and longyan, which were potentially useful for the treatment of Alzheimer’s disease. At the same time, the network-based methods of AI have also been successfully used for herb-target prediction. For example, Vanunu O et al. proposed the PRIoritizatioN and Complex Elucidation (PRINCE) method [111], which was originally used in disease-gene prediction and then was introduced into TCM target prediction by Yang et al. [112]. In 2019, Wang N et al. [113] presented an Herb-Target Interaction Network (HTINet) approach, a novel network integration pipeline for herb-target prediction mainly relying on symptom-related associations, which further reflected the increasingly prominent advantages of AI in network pharmacology.

AI also provides an alternative approach to the development of new drugs in traditional Ayurveda medicine. Ayurvedic medicine uses extracts from traditional medicinal plants to treat diseases [114]. Using a computational (in silico) approach, Fauzi [115] predicted potential targets for Ayurvedic anticancer compounds, which were obtained from an Indian plant anticancer database, with their chemical structures. The compounds can potentially be developed into potential new drugs.

Biomechanisms of Syndrome

Syndrome is the main type of clinical diagnosis for TCM. Through a comprehensive analysis of clinical information obtained through the four diagnostic TCM procedures, the nature of the disease is revealed and the syndrome type is classified [116]. AI methods have been integrated with system biology as a new route for the investigation of the underlying biological mechanisms of TCM syndrome, which has become a popular research topic.

Studies related to the biological mechanisms of syndrome have been carried out since 1950s and have made a number of significant research achievements [117], for example, the kidney Yang deficiency syndrome [118, 119], blood stasis syndrome, and spleen deficiency syndrome [120], which mainly relied on the related gene expressions and gene polymorphisms of model animals. However, these studies are still restricted by a bottleneck of problems, especially in the difference between an animal model and a human body. In 2004, related integrative text mining and network medicine studies compiled about two and half million disease-relevant PubMed citations and 1,479,630 human-gene-relevant PubMed citations in a local database. Bootstrapping techniques were used to extract Chinese disease names from the literature, and term cooccurrence was used to extract the relationships. Finally, 200,000 syndrome-gene relations were found by related shared diseases on MEDLINE [129], and syndrome-based gene networks were constructed to analyze the functional knowledge of genes from a syndrome perspective [130]. Since the publication of the classic work on disease omics in 2007

[121], network medical research on the analysis of disease mechanism by molecular interaction networks at various levels has become a hot topic [122]. Therefore, the research on the biological basis of syndrome from the perspective of molecular network has become a hot direction. The typical studies include the work from Li's group on cold and heat syndrome in 2007 [123, 124]. Furthermore, through the clinical detection of typical patients with cold and heat syndrome of gastritis, it is found that patients with cold and heat syndrome have the characteristics of imbalance of energy metabolism-immune regulation network, and the key nodes of the network can be used as potential biomarkers of cold and heat syndrome [125]. In addition, there are many studies of syndromes related to chronic diseases (such as type 2 diabetes and Qi-Yin deficiency syndrome [126], chronic kidney disease and blood stasis syndrome [127], asthma, and heat syndrome [128]). Several researchers in the field of TCM have further proposed new research ideas, such as syndrome omics [131], syndrome system biology, and network syndrome science [132, 133] to investigate the biological mechanisms of syndrome. To address the bottleneck of genetic associations of syndrome, Yang et al. curated a benchmark data set of 18,270 symptom-gene associations between 505 symptoms and 4549 genes and developed a deep-learning-based symptom candidate gene prediction algorithm [134], which have the potential to promote investigations of the molecular mechanisms of symptoms and thus provide candidate genes of syndrome (through syndrome-symptom connection) for validation in experimental settings. Obviously, the use of data mining techniques and bioinformatics approaches, the biological basis of TCM syndrome and symptom, would be further clarified in the future.

Shortcomings and Future Directions

Although AI has been used in AM to solve many practical medical problems, there still exist some obvious shortcomings compared with those of AI in modern biomedicine. Here we list several main aspects.

First, lack of high-quality data sets and shared terminological systems—although the application of AM becomes even wider to new arenas, such as telemedicine, mobile medicine, and Internet medicine [135], there is currently no substantial data sets for automatic diagnosis and treatment. Furthermore, terminological systems and ontologies that shared in the fields to help address the issues of semantic interchanges and communications are rare, which would make it difficult to accurately represent the data derived from various real-world clinical sources [23], especially in terms of symptom terminology, which traditional medicine is concerned about. For example, some patients have the sign of “limb pain,” while others have the sign of “lower limb pain.” It is clear that the information of “limb pain” covers the information of “lower limb pain” [136]. Thus, it is impossible to collect high-quality structured data if comprehensive standard clinical terminologies are not available. And more importantly, there is still lack of a knowledge graph that combines with AM and modern medicine, though the link between basic medical research and clinical treatment is crucial to the development of AM.

Second, lack of a comprehensive decision support system—although there are a number of mature systems for diagnosis prediction and prescription recommendation [71, 72], it is difficult to claim precision when it comes to personalized medicine. In fact, more and more populations actually use or are interested in AM services, such as people from South Korea, Australia (in 2005, a survey suggested that 68.9% of participants used at least one of the 17 forms of complementary medicine, <https://www.who.int/westernpacific/health-topics/traditional-complementary-and-integrative-medicine>), and the United Kingdom [137]. Herbalism, aromatherapy, homoeopathy, acupuncture, massage, and reflexology are among the most popular. However, as they become more popular, debates over the efficacy of AM have been highly polarized. Some indicate that modern medicine therapy is mostly practiced by medically trained doctors. It is not the same with AM as, for example, yoga and massage are mostly performed by patients at home [138]. Currently, information on the effectiveness of treatment is

often absent in most real-world TCM clinical data [136]. Therefore, most related TCM clinical data mining tasks only generate empirical regularities with no constraints on clinical effectiveness. This means that it is difficult to judge which discovered regularities and decision support are effective for disease treatment, in particular individualized sequential treatment decision policies.

Third, lack of substantial AI applications in AM research other than TCM—compared with TCM, there is much less research on other AM therapies (e.g., mind-body therapies), most of which focus on the treatment and safety aspect [139]. For example, a massage therapy for musculoskeletal pain can be very relaxing to the mind and relieve psychological distress [140]; mind and body practices may be of benefit in reducing stress; and hypnosis has consistently shown to have a clinically significant effect on reducing hot flashes [140]. A large amount of higher quality studies hope to determine efficacy and safety among all reviewed AM interventions for symptoms [3]. However, with the availability of large-scale big healthcare data, AM needs to give full play to the advantages that would better serve patients and society through the use of AI technology. The future direction of AI research requires multidisciplinary collaboration between different professionals, researchers, and scientists, as well as collaborative clinical practice with expertise from biomedical, bioinformatics and pharmaceutical disciplines.

Conclusion

The curation of big data makes AI technology and AM fields much closer and their development more closely interconnected in multiple clinical applications, such as intelligent diagnosis, treatment recommendation, drug discovery, and biological mechanisms of syndrome. Besides of the ontology and knowledge graph development, related AI methodologies in AM fields have transferred from production rules to data mining and now to deep learning, including applications for EMR information extraction, image-based clinical diagnosis, and integrative mining of both

clinical and biological data. Alternative medicine proposes a typical personalized medicine solution for chronic disease management, which involves various types of individualized data and biomedical research data. Many substantial challenges (e.g., data privacy, high-quality clinical data, large-scale knowledge graph, human-machine interaction) should be addressed or developed to improve the usability and actual benefits for AM practitioners. However, the large-scale real-world AM clinical practice with essential effectiveness provides a promising future and arena for both the application and research of AI in AM field to promote AM practice from empirical experiences to evidence-based personalized medicine.

References

- Aakster C. Concepts in alternative medicine. *Soc Sci Med.* 1986;22(2):265–73.
- Veizari Y, Leach MJ, Kumar S. Barriers to the conduct and application of research in complementary and alternative medicine: a systematic review. *BMC Complement Altern Med.* 2017;17(1):1–14.
- Steinhorn DM, Din J, Johnson A. Healing, spirituality and integrative medicine. *Ann Palliat Med.* 2017;6(3):237–47.
- Bao Y, et al. Complementary and alternative medicine for cancer pain: an overview of systematic reviews. *Evid Based Complement Alternat Med.* 2014;2014:170396.
- Murat-Ringot A, Preau M, Piriou V. Médecines alternatives complémentaires en cancérologie et essais randomisés [Complementary and alternative medicine in cancer patients and randomized controlled trials]. *Bull Cancer.* 2021;108(1):102–16.
- Deng G. Integrative medicine therapies for pain management in cancer patients. *Cancer J (Sudbury, Mass.).* 2019;25(5):343.
- Qi F, et al. The advantages of using traditional Chinese medicine as an adjunctive therapy in the whole course of cancer treatment instead of only terminal stage of cancer. *Biosci Trends.* 2015;9(1):16–34.
- Lü S, et al. The treatment of rheumatoid arthritis using Chinese medicinal plants: from pharmacology to potential molecular mechanisms. *J Ethnopharmacol.* 2015;176:177–206.
- Zhou L, et al. Systematic review and meta-analysis of traditional Chinese medicine in the treatment of migraines. *Am J Chin Med.* 2013;41(05):1011–25.
- Xiao L, Tao R. Traditional Chinese medicine (TCM) therapy. In: Substance and non-substance addiction. Springer Singapore; 2017. p. 261–80.
- Xiong X, et al. Chinese herbal medicine for coronavirus disease 2019: a systematic review and meta-analysis. *Pharmacol Res.* 2020;160:105056.
- Shu Z, et al. Clinical features and the traditional Chinese medicine therapeutic characteristics of 293 COVID-19 inpatient cases. *Front Med.* 2020;14:760–75.
- Shu Z, et al. Add-On Chinese Medicine for Coronavirus Disease 2019 (ACCORD): A Retrospective Cohort Study of Hospital Registries. *Am J Chin Med.* 2021;49(3):543–575.
- Evans RS, Pestotnik SL. Applications of medical informatics in antibiotic therapy. In: Antimicrobial susceptibility testing. Springer; 1994. p. 87–96.
- Ngiam KY, Khor W. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol.* 2019;20(5):e262–73.
- Ramesh A, et al. Artificial intelligence in medicine. *Ann R Coll Surg Engl.* 2004;86(5):334.
- Rauschert S, et al. Machine learning and clinical epigenetics: a review of challenges for diagnosis and classification. *Clin Epigenetics.* 2020;12:1–11.
- Nogales A, et al. A survey of deep learning models in medical therapeutic areas. *Artif Intell Med.* 2021;112(12):102020.
- Stanfill MH, Marc DT. Health information management: implications of artificial intelligence on healthcare data and information management. *Yearb Med Inform.* 2019;28(1):56.
- Tiwari P, et al. Recapitulation of Ayurveda constitution types by machine learning of phenotypic traits. *PLoS One.* 2017;12(10):e0185380.
- Zhou X, et al. Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support. *Artif Intell Med.* 2010;48(2–3):139–52.
- Feng Y, et al. Knowledge discovery in traditional Chinese medicine: state of the art and perspectives. *Artif Intell Med.* 2006;38(3):219–36.
- Zhou X, Peng Y, Liu B. Text mining for traditional Chinese medical knowledge discovery: a survey. *J Biomed Inform.* 2010;43(4):650–60.
- Ru J, et al. TCMSP: a database of systems pharmacology for drug discovery from herbal medicines. *J Chem.* 2014;6(1):1–6.
- Xu H, et al. ETCM: an encyclopaedia of traditional Chinese medicine. *Nucleic Acids Res.* 2019;47(D1):D976–82.
- Fang S, et al. HERB: a high-throughput experiment-and reference-guided database of traditional Chinese medicine. *Nucleic Acids Res.* 2021;49(D1):D1197–206.
- Wu Y, et al. SymMap: an integrative database of traditional Chinese medicine enhanced by symptom mapping. *Nucleic Acids Res.* 2019;47(D1):D1110–7.
- Zhou X, et al. Ontology development for unified traditional Chinese medical language system. *Artif Intell Med.* 2004;32(1):15–27.
- Guo Y, et al. Preliminary study on the characteristic elements of TCM clinical terminology standardization based on SNOMED CT core framework. *Chin J TCM Inf.* 2008;09:96–7. (in Chinese).
- Studer R, Benjamins VR, Fensel D. Knowledge engineering: principles and methods. *Data Knowl Eng.* 1998;25(1–2):161–97.

31. Zong X, Dai L. Analysis of 196 cases of liver disease treated by computer. *Liaoning J Tradit Chin Med.* 1992;06:26–7. (in Chinese).
32. Tian H, Zhou G. Design and implementation of GTS model for general Chinese medicine expert system. *Acta Comput Sin.* 1987;08:508–12. (in Chinese).
33. Luo Y. Implementation method and technology of monkey, an expert system of traditional Chinese medicine. *Acta Comput Sin.* 1988;(06):371–7. (in Chinese).
34. Jin Z, Liu F, Yu X. Chinese medicine expert system tool YHW-CTMEST. *Comput Eng Appl.* 1988;06: 59–63. (in Chinese).
35. Nopparatkit P, Nagara B, Chansa-ngavej C. Expert system for skin problem consultation in Thai traditional medicine. *Afr J Tradit Complement Altern Med.* 2014;11(1):103–8.
36. Shojaee-Mend H, Ayatollahi H, Abdolahadi A. Development and evaluation of ontologies in traditional medicine: a review study. *Methods Inf Med.* 2019;58(06):194–204.
37. Long H, et al. An ontological framework for the formalization, organization and usage of TCM-knowledge. *BMC Med Inform Decis Mak.* 2019;19(2):79–89.
38. Yu T, et al. Research on the construction of large-scale TCM knowledge map. *Chin Digit Med.* 2015;10(03): 80–3. (in Chinese).
39. Mohan AR, Arumugam G. Developing Indian medicinal plant ontology using OWL and SWRL. In: International conference on data engineering and management. Springer; 2010.
40. Jang H, et al. Ontology for medicinal materials based on traditional Korean medicine. *Bioinformatics.* 2010;26(18):2359–60.
41. Kouame A, et al. Visual representation of African traditional medicine recipes using icons and a formal ontology, onto MEDTRAD. *Stud Health Technol Inform.* 2020;270:791–5.
42. Wang X, et al. A self-learning expert system for diagnosis in traditional Chinese medicine. *Expert Syst Appl.* 2004;26(4):557–66.
43. Cortes C, Vapnik V. Support-vector networks. *Mach Learn.* 1995;20(3):273–97.
44. You M, et al. MAPLSC: a novel multi-class classifier for medical diagnosis. *Int J Data Min Bioinform.* 2011;5(4):383–401.
45. Rish I. An empirical study of the naive Bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence. 2001.
46. Peterson LE. K-nearest neighbor. *Scholarpedia.* 2009;4(2):1883.
47. Kanawong R, et al. Automated Tongue Feature Extraction for ZHENG Classification in Traditional Chinese Medicine. *Evid Based Complement Alternat Med.* 2012;2012:912852.
48. Qi Z, et al. The Classification of Tongue Colors with Standardized Acquisition and ICC Profile Correction in Traditional Chinese Medicine. *Biomed Res Int.* 2016;2016:3510807.
49. Zhang J, et al. Diagnostic Method of Diabetes Based on Support Vector Machine and Tongue Images. *Biomed Res Int.* 2017;2017:7961494.
50. Zhao C, et al. Advances in Patient Classification for Traditional Chinese Medicine: A Machine Learning Perspective. *Evid Based Complement Alternat Med.* 2015;2015:376716.
51. Li F, et al. Computer-assisted lip diagnosis on traditional Chinese medicine using multi-class support vector machines. *BMC Complement Alternat Med.* 2012;12(1):1–13.
52. Yao M, et al. Analysis of the association rule in the composition of the TCM formulas for diabetes. *J Beijing Univ TCM.* 2002;25(6):48–50. (in Chinese).
53. Huang K, et al. Law of acupoint selection in acupuncture treatment for insomnia based on data mining method. *Chin Acupunct Moxibustion.* 2015;35(09): 960–3. (in Chinese).
54. Yongjian L, Zhaoqin F. Application of clustering analysis in TCM research. *J Nanjing Univ Tradit Chin Med.* 2001;03:182–4. (in Chinese).
55. Qi M, Luo X, Wang X. Fuzzy cluster analysis of *Scutellaria baicalensis* varieties and substitutes. *J Shenyang Coll Pharm.* 1992;02:127–9. (in Chinese).
56. Wang F, et al. Cluster analysis on traditional Chinese medicine syndrome of metaphase and advanced lung cancer. *Chin J Inf TCM.* 2006;10:28–9. (in Chinese).
57. Tang S, et al. Designing new TCM prescriptions based on complex system entropy cluster. *World Sci Technol/Mod Tradit Chin Med Mater Med.* 2009;11 (02):225–8. (in Chinese).
58. Mourad R, et al. A survey on latent tree models and applications. *J Artif Intell Res.* 2013;47:157–203.
59. Zhang NL, et al. Latent tree models and diagnosis in traditional Chinese medicine. *Artif Intell Med.* 2008;42(3):229–45.
60. Cao C, Wang H, Sui Y. Knowledge modeling and acquisition of traditional Chinese herbal drugs and formulae from text. *Artif Intell Med.* 2004;32(1): 3–13.
61. Liu GP, et al. Deep learning based syndrome diagnosis of chronic gastritis. *Comput Math Methods Med.* 2014;2014:938350.
62. Yeh H-Y, et al. Predicting the associations between meridians and Chinese traditional medicine using a cost-sensitive graph convolutional neural network. *Int J Environ Res Public Health.* 2020;17(3):740.
63. Ji B, et al. Research on Chinese medical named entity recognition based on collaborative cooperation of multiple neural network models. *J Biomed Inform.* 2020;104:103395.
64. Shu Z, et al. Symptom-based network classification identifies distinct clinical subgroups of liver diseases with common molecular pathways. *Comput Methods Prog Biomed.* 2019;174:41–50.
65. Li W, Yao X. Application of data mining technology in field of pediatric massage. *Acta Chin Med.* 2019;34 (06):1193–6. (in Chinese).
66. Ai M, A 7-joint robot that can play Taijiquan. *Modern manufacturing.* 2016(26):62–63. (in Chinese).
67. Daynac M, Cortes-Cabrera A, Prieto JM. Application of Artificial Intelligence to the Prediction of the Antimicrobial Activity of Essential Oils. *Evid Based Complement Alternat Med.* 2015;2015:561024.

68. Tang Y, et al. Classification of tongue image based on multi-task deep convolutional neural network. *Comput Sci.* 2018;45(12):255–61. (in Chinese).
69. Huang Q. Research on the auxiliary system of pulse diagnosis in traditional Chinese medicine based on artificial intelligence. *Shanxi Univ Sci Technol.* 2018;2018(03):33. (in Chinese).
70. Cui J. The immersive acupuncture training system based on machine learning. Graduate School of Tianjin University; 2019. (in Chinese).
71. Yin T, et al. Progress and prospect: acupuncture efficacy of combining machine learning with neuroimaging properties. *World Chin Med.* 2020;15(11):1551–4. (in Chinese).
72. Qian H. Research and implementation of question answering system based traditional Chinese medicine semantic web. Zhejiang University; 2016. (in Chinese).
73. Yuan N, et al. Depth representation-based named entity extraction for symptom phenotype of TCM medical record. *World Sci Technol-Mod Tradit Chin Med.* 2018;20(3):255–362. (in Chinese).
74. Wang N, et al. Application of artificial intelligence-assisted diagnostic system in diagnosis and treatment of coronavirus disease 2019. *China Med Dev.* 2020;35(6):75–79. (in Chinese).
75. Wang Y, et al. Tongue image color recognition in traditional Chinese medicine. *J Biomed Eng.* 2005;22(6):1116–20. (in Chinese).
76. Feng Y, et al. Modeling method and validation research of partially observable Markov decision-making process model in the optimization of clinical treatment for unstable angina pectoris with integrated traditional Chinese and western medicine. *Chin Gen Pract.* 2020;23(17):2181–5. (in Chinese).
77. Pang B, et al. Cognitive model for diagnosis and treatment of lung cancer based on the prototype category theory optimization ideas and methods. *Beijing J Tradit Chin Med.* 2018;37(12):1141–5. (in Chinese).
78. Zhang L, et al. Anxiety and depression in patients with polycystic ovary syndrome and relevant factors. *J Int Reprod Health/Fam Plann.* 2018;37(4):269–72. (in Chinese).
79. Chen J, et al. Development of clinical assistant diagnosis and treatment system based on traditional Chinese medicine syndrome differentiation and treatment. *World Sci Technol-Mod Tradit Chin Med.* 2015;017(012):2436–42. (in Chinese).
80. Li J, Hu J, Wang Y. Research on establishment of standard model of TCM syndrome diagnosis based on data mining of type 2 diabetes. *Chin J Basic Med Tradit Chin Med.* 2008;05:367–70. (in Chinese).
81. Joshi A, et al. Arterial pulse system: modern methods for traditional Indian medicine. In: 2007 29th annual international conference of the IEEE engineering in medicine and biology society. IEEE; 2007.
82. Kim K-B, Kim J-W. Self health diagnosis system with Korean traditional medicine using fuzzy art and fuzzy inference rules. In: Asian conference on intelligent information and database systems. Springer; 2012.
83. Ying C, Lizhuang M, Jiatuo X. Analysis on the research status of TCM syndrome differentiation information. *Liaoning J Tradit Chin Med.* 2009;36(03):486–8. (in Chinese).
84. Guo R. Computer medical diagnosis system. *Appl Electron Techn.* 1982;(06):7–9+23. (in Chinese).
85. Liu X, et al. Research and practice of intelligent analysis system of Chinese medicine prescription. *Chin J Inf Tradit Chin Med.* 2007;10:97–9. (in Chinese).
86. Sun X, et al. Technology and application of TCM clinical assistant decision system. *New Era Technol.* 2017;04:56–60. (in Chinese).
87. Ruan C, et al. THCluster: herb supplements categorization for precision traditional Chinese medicine. In: 2017 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE; 2017.
88. Liu J, Jiang W, Shen G. Design of TCM intelligent expert system for hypertension based on data analysis. *Beijing J Tradit Chin Med.* 2019;38(09):904–906+945. (in Chinese).
89. Ikram RRR, Abd Ghani MK, Abdullah N. An analysis of application of health informatics in Traditional Medicine: a review of four Traditional Medicine Systems. *Int J Med Inform.* 2015;84(11):988–96.
90. Janmejaya S. Advancements in Indian System of Medicine (ISM) informatics: an overview. *Glob J Res Med Plants Indigenous Med.* 2013;2(7):546.
91. Yin T, et al. Progress and prospect of machine learning in research of acupuncture and moxibustion. *Chin Acupunct Moxibustion.* 2020;40(12):1383–6. (in Chinese).
92. Lam CFD, et al. Chinese acupuncture expert system (CAES) – a useful tool to practice and learn medical acupuncture. *J Med Syst.* 2012;36(3):1883–90.
93. Wen C, Yuan X, Wang J. Design and development of clinical acupuncture auxiliary software and hardware integration system. *Shanghai J Acupunct Moxibustion.* 2012;31(12):930–1. (in Chinese).
94. The First Chinese Medicine Master Cheng Shennong Meridian Robot System Was Launched. *TCM Healthy Life-Nurturing.* 2017. 000(010): 2–2. (in Chinese).
95. Zhang J, et al. Discussion on the research and development of digital meridian intelligent acupuncture robot. *Guid J Tradit Chin Med Pharm.* 2018;24(19):66–8. (in Chinese).
96. Lan K-C, Litscher G. Robot-controlled acupuncture – an innovative step towards modernization of the ancient traditional medical treatment method. *Medicines.* 2019;6(3):87.
97. Litscher G. Robot-assisted acupuncture – a technology for the 21st century. *Acupunct Auricular Med.* 2017;4:9–10.
98. Kinkhabwala D, Bhavesh A. Can meditation practices be elevated, for the higher level of consciousness, taking help of artificial intelligence? Taking Help of Artificial Intelligence; 2020.
99. Vijayaragavan GR, et al. EEG monitored mind de-stressing smart phone application using Yoga and

- Music Therapy. In: 2015 international conference on green computing and internet of things (ICGCIoT). IEEE; 2015.
100. Kosunen I, et al. RelaWorld: neuroadaptive and immersive virtual reality meditation system. In: Proceedings of the 21st international conference on intelligent user interfaces. 2016.
101. Prasanna M, Arunkumar T, Arunkumar S. A Mobile application based smart system for supporting yoga activities and health monitoring. *Res J Pharm Technol.* 2017;10(11):3863–7.
102. Li S, Bo Z. Traditional Chinese medicine network pharmacology: theory, methodology and application. *Chin J Nat Med.* 2013;11(2):110–20.
103. Hopkins AL. Network pharmacology. *Nat Biotechnol.* 2007;25(10):1110–1.
104. Li S. Framework and practice of network-based studies for Chinese herbal formula. *J Chin Integr Med.* 2007;5(5):489–93.
105. Hopkins AL. Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol.* 2008;4(11):682–90.
106. Li S, Zhang B, Zhang N. Network target for screening synergistic drug combinations with application to traditional Chinese medicine. *BMC Syst Biol.* 2011;5(1):1–13.
107. Yu G, et al. Network pharmacology-based identification of key pharmacological pathways of Yin-Huang-Qing-Fei capsule acting on chronic bronchitis. *Int J Chron Obstruct Pulmon Dis.* 2017;12:85.
108. Wang Y, et al. Predicting new indications of compounds with a network pharmacology approach: Liuwei Dihuang Wan as a case study. *Oncotarget.* 2017;8(55):93957.
109. Ma X, et al. Study on the biological basis of hypertension and syndrome with liver-fire hyperactivity based on data mining technology. *World J Tradit Chin Med.* 2018;4(4):176.
110. Fang J, et al. Network pharmacology-based study on the mechanism of action for herbal medicines in Alzheimer treatment. *J Ethnopharmacol.* 2017;196:281–92.
111. Vanunu O, et al. Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol.* 2010;6(1):e1000641.
112. Yang K, et al. Integrating herb effect similarity for network-based herb target prediction. In: 2015 8th international conference on biomedical engineering and informatics (BMEI). IEEE; 2015.
113. Wang N, et al. Herb target prediction based on representation learning of symptom related heterogeneous network. *Comput Struct Biotechnol J.* 2019;17:282–90.
114. Corson TW, Crews CM. Molecular understanding and modern application of traditional medicines: triumphs and trials. *Cell.* 2007;130(5):769–74.
115. Fauzi FM, et al. Linking Ayurveda and Western medicine by integrative analysis. *J Ayurveda Integr Med.* 2013;4(2):117.
116. Su SB, et al. Evidence-based ZHENG: a traditional Chinese medicine syndrome. Hindawi; 2012.
117. Ai-ping L. Research on the combination of disease and syndrome in animal models: from theoretical innovation to technical challenges. *Chin J Integr Tradit West Med.* 2013;01:6–7. (in Chinese).
118. Shen Z. Study on the localization of kidney-yang deficiency syndrome. *Chin J Integr Tradit West Med.* 1997;1:50–2. (in Chinese).
119. Jiang Y. Basic research on blood stasis syndrome. *Chin J Basic Med Tradit Chin Med.* 2005;8:561–3. (in Chinese).
120. Kong L. Current situation and prospect of TCM spleen deficiency syndrome research. *Beijing J Tradit Chin Med.* 2008;09:738–9. (in Chinese).
121. Goh K-I, et al. The human disease network. *Proc Natl Acad Sci.* 2007;104(21):8685–90.
122. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet.* 2011;12(1):56–68.
123. Li S, et al. Understanding ZHENG in traditional Chinese medicine in the context of neuro-endocrine-immune network. *IET Syst Biol.* 2007;1(1):51–60.
124. Chen G, et al. A network-based analysis of traditional Chinese medicine cold and hot patterns in rheumatoid arthritis. *Complement Ther Med.* 2012;20(1–2):23–30.
125. Li R, et al. Imbalanced network biomarkers for traditional Chinese medicine syndrome in gastritis patients. *Sci Rep.* 2013;3(1):1–7.
126. Xie Y, et al. Association of APOE polymorphisms and insulin resistance with TCM syndromes in type 2 diabetes patients with macroangiopathy. *Mol Med Rep.* 2011;4(6):1219–23.
127. Li S, et al. Study on the relationship between blood stasis syndrome and clinical pathology in 227 patients with primary glomerular disease. *Chin J Integr Med.* 2009;15(3):170–6.
128. Hsu CH, et al. High eosinophil cationic protein level in asthmatic patients with “heat” Zheng. *Am J Chin Med.* 2003;31(02):277–83.
129. Wu Z, et al. Text mining for finding functional community of related genes using TCM knowledge. In: European conference on principles of data mining and knowledge discovery. Springer; 2004.
130. Zhou X, et al. Integrative mining of traditional Chinese medicine literature and MEDLINE for functional gene networks. *Artif Intell Med.* 2007;41(2):87–104.
131. Wang Z, Wang J, Wang Y. The study of TCM syndromes in the post-genomic era. *Chin J Integr Tradit West Med.* 2001;08:621–3. (in Chinese).
132. Shen Z. Systems biology and the study of TCM syndromes. *Chin J Integr Tradit West Med.* 2005;03:255–8. (in Chinese).
133. Li Z. Network syndrome: a new model of TCM syndrome research. *Tianjin Tradit Chin Med.* 2015;07:388–92. (in Chinese).
134. Yang K, et al. Heterogeneous network embedding for identifying symptom candidate genes. *J Am Med Inform Assoc.* 2018;25(11):1452–9.

135. Hu NZ, et al. A cloud system for mobile medical services of traditional Chinese medicine. *J Med Syst.* 2013;37(6):1–13.
136. Liu B, et al. Data processing and analysis in real-world traditional Chinese medicine clinical data: challenges and approaches. *Stat Med.* 2012;31(7):653–60.
137. Ernst E. The role of complementary and alternative medicine. *BMJ.* 2000;321(7269):1133.
138. Field T, et al. Yoga and massage therapy reduce prenatal depression and prematurity. *J Bodyw Mov Ther.* 2012;16(2):204–9.
139. Tang SW, Tang WH, Leonard BE. Safety of traditional medicine and natural product supplements in psychiatry. *Int Clin Psychopharmacol.* 2020;35(1):1–7.
140. Rapaport MH, et al. Massage therapy for psychiatric disorders. *Focus.* 2018;16(1):24–31.



Umar Iqbal and Junaid Nabi

Contents

Introduction	1263
AI Tools	1264
Cancer Detection	1265
Cancer Treatment	1268
Conclusion	1270
References	1270

Abstract

While the survival rate for cancer patients has improved from 49% in 1970 to 70% in 2020, much of this progress has been attributed to improvements in and focus on early detection. With numerous advances in cancer therapy – from leveraging genomics to immunology – the overall consensus on progress has been underwhelming. With the limited availability of cancer care specialists – even for well-resourced and high-income countries – Artificial Intelligence (AI) has the potential to fill this void and accelerate development of diagnostic and treatment options, formulation of

therapeutic plans, delivery of the treatment, monitoring, and surveillance. We describe how artificial intelligence techniques – especially machine learning – can be applied to clinical data from radiology, pathology, and electronic health records to supplement – and often optimize and automate – clinical decision making.

Keywords

Artificial intelligence · Artificial intelligence in oncology · Machine learning algorithms · Precision oncology · Automated decision-making systems · Supervised learning · Training sets in oncology

Introduction

English mathematician Charles Babbage conceived the idea of a steam-driven “Analytical Engine” in 1822 for computing tables of numbers.

U. Iqbal
ATLAS Program, Department of Urology, Roswell Park Cancer Institute, Buffalo, NY, USA

J. Nabi (✉)
Harvard University, Boston, MA, USA

This rudimentary machine was arguably the first one to demonstrate that machines can reliably follow, interpret, and apply logic. The question of whether machines can go beyond logic and enter the realm of “thinking” was initially conceived by Alan Turing in his seminal paper “Computing Machinery and Intelligence,” published in *Mind*, October 1950 [1]. The affirmative tone of the paper was popular and positively adopted by the wider academia. This led to substantial overselling of its practical potential that was not in sync with the available computing power of the day. Thereafter, artificial intelligence (AI)-related research entered the so-called AI winter [2]. However, the first two decades of the twenty-first century have witnessed an exponential rise of AI-powered technologies – often owing to multiple factors, such as availability of big data, cloud computing, advances in machine learning, increased investment, and immense interest from industry [3].

Application of AI to problems of health care are of particular interest because of inadequate manpower, disparate training, lack of specialists, resource crunch, and diagnostic and therapeutic errors [4]. Currently, role of AI is being explored for diagnostics (in specialties such as radiology, histology, and dermatology), therapeutics (including personalized medicine and surgical planning), rehabilitation (monitoring, biofeedback), prognostication, and research. Within the healthcare setting, it appears that oncological practice is uniquely poised to benefit from advances in AI. Global population is aging – an important contributor to increased cancer incidence. Oncological care is nearly always multidisciplinary with significant collaboration among various interrelated medical specialties. Adding to this issue is a projected shortage of cancer care providers in the immediate future, even for well-resourced and high-income countries [5]. AI can mitigate the impact of these changes by assisting clinicians in making better decisions, share workload, prevent burnout, and eventually improve patient outcomes. With this in mind, we will briefly introduce the tools for AI, current status of AI in oncology, future directions, and possible limitations.

AI Tools

The central theme of AI is to enable technologies or devices to perform tasks that normally require human intellect or capability. AI approach employs complex mathematical modeling and computer science to approximate the function of human neurons [6]. Human intellect is a very sophisticated and evolved system that utilizes the sensory system to gather data and then employ cognition, perception, logic, pattern detection, and linguistics to make sense of it. The end goal is to respond to the environment in a meaningful and beneficial manner. AI seeks to replicate this process by equipping machines with means to ingest data via cameras, sensors, historical clinical data, and give them the ability to continuously learn and refine their response to eventually mimic human understanding and actions. The armament of AI in degrees of increasing complexity is briefly discussed here:

- **Predictive Analytics (PA)** [7]: As the name suggests, PA utilizes current and historical data, data mining, conventional statistics, and machine learning – either alone or in combinations to make predictions about future events. The most well-known and oft-cited example is that of credit scoring. Banking history like payments, loans, consumer data, and current financial standing is used to predict the future likelihood of making credit payments on time. Social media algorithms determining your suggested YouTube videos and news feed on Facebook are other pertinent examples of predictive analytics. In medicine, a patient’s health records and those of a similar population cohort can be used to predict disease and treatment outcomes.
- **Computer Vision (CV)** [8]: Human beings are quite adept at analyzing visuals both still and in motion. CV is a rapidly advancing field that aims to equip computers with a human-like ability to parse through images and videos to make meaningful conclusions and respond to them. The applications range from the humble barcode-based sorting of parcels to automatic cars on the highways. In health care, CV is

- being extensively studied to aid in development of automated means of interpreting radiological scans, histological images, identifying skin lesions, and assessment of provider skills.
- **Natural Language Processing [NLP] [9]:** The fusion of linguistics and computer science to comprehend human speech in both text and spoken form as well as de novo synthesis of speech encompasses the term NLP. It is important to emphasize here that “natural” refers to everyday syntax, emotions, sarcasm, and slang. Ubiquitous virtual assistants like *Siri* and *Alexa* are good examples of how far this field has come. The applications and possibilities in health care are immense. NLP can be used to gather relevant data from patient notes that are already in the electronic health records (EHR), perform scribe duties, and possibly gather information about patient wellbeing. NLP has been used to identify markers of psychiatric symptoms and suicidal ideation in patients [10]. This may also be used to alert the system about provider fatigue and wellbeing.
 - **Machine Learning [11, 12]:** This is a subset of AI that provides computer algorithms the automated ability to improve with experience. The basic algorithm is fed training data, and it learns and improves continuously with subsequent data sets. This is especially important for complicated tasks where it is virtually impossible to code for the algorithms as innumerable possibilities exist. It makes more sense to let the machine decipher patterns, learn on their own, and refine the algorithm. The earliest examples of machine learning can be found in web-based search engines, spam filters for emails, online fraud detection, etc. ML in health care is all encompassing and can provide the structure for all the tools discussed previously (PA, CV, NLP)
 - **Artificial Neural Networks and Deep Learning [ANN and DL] [13]:** ANNs are computational systems that aim to mimic the neuronal circuitry and decision making of the animal or human brain. Just like numerous layers of neurons in the brain, ANNs have layers of nodes. Each layer extricates relevant variables from

the input data and passes it on to the next level usually in increasing orders of complexity. The variables may be weighted and passed on only when they reach a certain threshold. After passing through multiple layers, the ANNs can identify hidden and abstract information and present a meaningful output. ANNs with enough depth and layers of complexity are termed as DL networks.

Cancer Detection

Probably the most exciting avenue for AI in health care is cancer detection. Early recognition of a possibly neoplastic process is usually the number one determinant of positive long-term outcome. There is significant disparity at the provider and institution level in the ability to recognize cancer early. Diagnosis of a neoplastic process is achieved either by radiological imaging, histology, genomics, metabolites, biomarkers, and their various combinations. AI can help cancer detection by analyzing radiographic, clinical and histological images, EHR data or even by discerning hitherto unknown patterns from the vast repository of medical data that is available digitally. AI for cancer detection can be leveraged both at the screening level and the expert level to augment existing health infrastructure.

- **AI for Image analysis:** Cancer detection and follow up can be performed on three major imaging subtypes (clinical, radiological, and histological). A brief introduction and relevant limitations are discussed below.
 1. **Radiological images [14]:** The use of data characterization algorithms to extract features from radiological images is termed as radiomics. Radiological images generated by magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography (PET), and ultrasonography (US) have all been used to extract radiomic features to diagnose or prognosticate in oncology [15]. Radiomics has clearly defined steps of image acquisition and preprocessing, tumor segmentation,

feature extraction, knowledge discovery, and modeling [16]. Unlike skin images, the advantage of radiomics is that even though there is variability due to image acquisition methods, the differences are mostly concrete and accounted for. Possible variables for CT (scanner, voxel, kernel), MRI (field strength, manufacturer, protocol), and PET (grid size, iterations, breathing artifacts) images have been previously described and their mitigators suggested [17–20]. Attempts to standardize data due to differences in scanner type, institutional protocols, and radiomics software used have been made at the pre-processing level [21]. After preprocessing, the next step is segmentation, that is, separation or marking the tumor that is termed “volume of interest.” This process can be manual, semi-automatic, or automatic. The non-geometric nature of the tumors, time constraints, and interoperator variability are leading the field toward automation for this step [21, 22]. Feature extraction is the core of radiomics and is achieved manually (semantic or non-semantic) and by DL methods, for example, the question of whether a lymph node harbors malignancy may use semantic features (shape, border), non-semantic (signal intensity, uniformity), and DL (convolutional features). DL features are superior to manual features as they also incorporate the clinical outcomes rather than just descriptive nature of manual features [16]. The last phase of knowledge discovery and modeling phase refers to choosing the relevant features (avoiding redundancy) and constructing a predictive model based on it. Models can be trained in supervised, semi-supervised, or unsupervised learning modes. Finally, the model that has been developed needs to be validated, preferably on an external cohort of images. At present even though multiple proof-of-concept studies have been performed, the data is usually from a single institution. This raises questions about reproducibility, generalization, and bias in

ML outputs [23]. Future studies will need to incorporate standardized protocols and may be assessed using the recently proposed radiomics quality score (RQS) [24].

2. **Clinical images** [25]: For the most part, these are limited to dermatological applications of AI. Much work has been done at the proof-of-concept level to leverage AI for differentiating benign versus malignant skin lesions, gradation of ulcers, and inflammatory skin conditions. In one of the landmark studies, Esteva et al. used Google Inception v3 convolutional neural network (CNN) architecture pretrained on the ImageNet dataset and fine-tuned on their own dataset to classify malignant melanomas versus benign nevi. They compared the performance of 21 board-certified dermatologists to their deep CNN and found that AI could classify the lesion at a level comparable to the dermatologists with area under the curves (AUC) a remarkable 0.94–0.96 [26]. Multiple other studies have shown similar promising results [27, 28]. While much of the research is focused on melanoma detection, there have been successful attempts at differentiating premalignant vs benign vs keratinocyte carcinomas as well [29, 30]. These results remain true both for direct clinical images and those at the dermoscopic level. Due to burgeoning interest in AI and dermatology there is an annual competition “International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge” [25]. The results so far have repeatedly demonstrated that AI can improve diagnostic accuracy for dermatologists [31, 32].

Dermatological image analysis has also been used for wound assessment of diabetic and pressure ulcers [33, 34]. Cancer patients are routinely bedridden for varying lengths of time owing to treatment or the disease itself and are prone to develop pressure ulcers. These tools, provided they gain clinical acceptance, can be valuable in prevention and treatment of these ulcers. AI in dermatology has particularly widespread

implications. With approximately six billion smartphones in play by the end of 2021, mobile-based resources can dramatically expand the diagnostic care to outside of the clinic [26].

However, as with every promising idea there are certain major limitations. The AI models are only as good as the input data. Biases pertaining to age, gender, race, and ethnicity may creep into the algorithms if the input data is lacking diversity [35]. Differences pertaining to race must be factored in the models as disease presentation can be wildly divergent due to difference in skin color. Indeed, Han et al. reported that the performance of DL algorithms can be enhanced if they were trained on multiple ethnic populations [36]. Proper input dataset can help preserve the generalizability of the algorithms. Also, the quality of the images must be standardized, and the lighting, angle, background, pen marking, etc. are all possible confounders [25]. Another possible limitation is the lack of haptic feedback, and inability to elicit clinical signs in the absence of a trained provider. And if we are to look toward a future where smartphones double up as diagnostic dermatology clinics, we need to be cognizant of the fact that the camera quality on the devices may differ. We should be careful that patients from higher socioeconomic strata may have better cameras on their smartphone, and consequently, the performance of AI is also better. This may inadvertently lead AI-based applications to have poorer accuracy in marginalized patient populations that should have been the target of any such innovations [23]. In summary, future work should be standardized, use large and diverse datasets, utilize EHR information, and focus on well-designed prospective controlled trials.

3. **Pathological images** [37]: While the initial focus of AI researchers was skewed toward radiologic images, now an analogous field of pathomics has also come into being. The underlying principles from data acquisition

to modeling are same as previously described in the radiomics section. With the widespread adoption of whole slide imaging, a vast repository of pathological images is generated daily [38]. AI can be used for tissue recognition, diagnosis, and even identifying disease biology. Diagnostically, AI has been shown to be capable of differentiating benign from malignant tissue, grading of cancers, and immunohistochemical analysis [39]. As a representative example, Alphabet Inc. (Google division) researchers developed an ANN that detected breast cancer metastases with an 89% accuracy rate compared to 73% for pathologists [40]. AI can especially be useful for cancers that are notorious for interobserver bias, for example, prostate cancer and its grading (Gleason score) [41]. ANNs have been developed that were able to outperform general pathologists for Gleason grading with an accuracy of 0.70 compared to 0.61 [42]. Demonstration of these and similar capabilities has led to proposals of AI being used as a triage system to improve workflow. On the first pass digital slides could be examined by AI algorithms and they flag potential malignant ones for review by the pathologist. The threshold could be set at a value such that sensitivity is not compromised. Alternately, AI could be used as the second opinion for relatively simple cases [37]. Importantly, most of the currently used AI algorithms use hematoxylin and eosin (H&E) stained sections and no additional preparation is required.

Identification of pathological features that are ordinarily not distinguishable by human eye is another interesting application. Simple H&E stained slides were used to predict SPOP mutation in prostate cancer, similarly BRAF mutations were recognized in melanomas [43, 44]. This has the potential to supplant IHC in the future [37]. Additionally, insights into tumor biology like aggressiveness have also been gleaned from simple H&E slides, for example, in

colorectal carcinoma [45]. Taking all these capabilities into consideration, AI has been shown to consistently perform at or above the level of pathologists for cancer detection [46].

As with other image-based AI applications, there is need for standardization. There is some concern about difference in color tone of the slides at the institutional level due to differing regents, protocols, and slice thickness [47]. The variability in these parameters needs to be minimized to improve AI performance and generalizability. Lastly, composite fields like radiohistomics and histogenomics may be fields of interest in the near future [37].

- **AI Utilizing EHR and Laboratory Data for Cancer Diagnosis**

While image analysis has been the core of cancer diagnosis using AI, it must be acknowledged that newer fronts are opening and deserve attention. AI has been used for cancer diagnosis by allowing it to trawl the vast repository of EHR data and discovering hitherto unknown patterns in it. Information contained in EHR, including patient history and physical examination, in combination with laboratory findings from biomarkers and genomics, have been shown to enable early cancer detection [48–50]. Zheng et al. used self-organizing map (SOM) models to identify metabolic markers that can aid in early diagnosis of renal cell carcinoma. The same model was also efficient in monitoring postoperative recovery [51]. A similar model for detection of bladder cancer has also been proposed [50].

Genetic mutations develop over years or sometimes decades, before eventually manifesting as a cancer phenotype [52]. The last decade has seen genome sequencing cost decrease substantially from USD 50,000 to roughly USD 1,000. This newfound affordability has given rise to the possibility of utilizing genomic data as a screening test or predictive test in high risk patients. Zhou et al. developed a DL system ExPecto that uses genomic data and links mutations with disease prediction [53]. Better understanding

and validation of these methods can aid in timely diagnosis.

Cancer Treatment

The passing of the US National Cancer Act in 1971 is regarded as a major springboard in the fight against cancer. The 5-year survival rate for cancer patients has gone up from 49% in 1970 to 70% in 2020. However, most of this progress is attributable to improvements and focus on early detection [54]. While there have been numerous pathbreaking advances in cancer treatment, for example, leveraging genomics and immunology, the consensus is that the overall progress has been underwhelming [55]. AI has the potential to fill this void and accelerate development of treatment options, formulation of treatment plans, delivery of the actual treatment, monitoring, and surveillance [56, 57]. Here, we briefly discuss the current status and future direction of AI in cancer treatment.

- **Clinical Decision making** – Somashekhar et al. compared AI generated (IBM Watson) recommendations with tumor boards and found 93% concordance for breast cancer [58]. Even more significantly, in a prospective series of 1000 cases clinicians changed their decision in 14% of cases after AI suggested solutions were presented [59]. The changes were made because AI identified newer treatment options (55%), better personalized alternative (30%), and genotypic/phenotypic data or evolving clinical experience (15%). The authors have suggested that AI will be an import complement rather than replacement of traditional decision-making apparatus. The pace of literature that is generated for oncology and allied fields is increasing every year. The proportion of cancer-related entries in PubMed has risen from 6% of all entries in 1950 to 16% in 2016 [60]. And with 200,000 cancer-related articles published in 2019 alone, it is safe to say that available human resources are inadequate to keep track of it [47]. AI can tap into this data and help truly herald the age of precision

medicine. As an example, let us take a patient with metastatic gastric carcinoma who has exhausted treatment options. A newly discovered mutation that is responsive to a known drug is identified and matched with the mutation present in this patient by AI in real time. Then this patient is immediately offered this drug or given the option of an appropriate clinical trial. This contrasts with percolation of information from article to average provider that can take anywhere between months to total ignorance. This is a lag that the patient cannot afford. The role of AI in genomics itself is also well defined [61]. Patel et al. showed that AI-based systems outperformed manual systems in identifying genomic variations of clinical significance [62]. These may be used to populate existing databases like Catalogue of Somatic Mutations in Cancer (COSMIC) that is presently curated by scientists. This registry links mutations with cancer development, drug response, and drug resistance [63].

AI has the potential to guide chemotherapy both in terms of choice and dosage. AI models have been shown to predict the sensitivity and resistance of tumors to chemotherapy [64]. CURATE.AI, an AI platform, was used by researchers in Singapore to guide the dosage of combination chemotherapy for metastatic castration resistant prostate carcinoma. There was enhanced efficacy and safety when dose modulations suggested by CURATE.AI were used [65]. Similar systems that are geared toward immunotherapy for cancer (e.g., programmed cell death protein-1) inhibitors have also been evaluated [66]. Lastly, AI systems also have the potential to prevent overtreatment by accurately identifying low-risk lesions, for example, in breast or cervix [67, 68].

- **Role in Surgery:** The interaction of AI and surgery can be broadly separated into preoperative, operative, and postoperative components [69]. AI-based accurate risk scoring systems may be developed that combine population, patient, surgeon, and institutional data with disease characteristics to provide an estimate of likely surgical outcome. During the preoperative phase, AI can be used for surgical

planning by using EHR data and imaging. Promising results have been reported in construction of realistic virtual or 3D models for defining surgical planes in complicated cases [70]. During the surgery itself, physiological data from various sensors monitor analyzed by AI can give real-time feedback on patient status. Lundberg et al. developed an explainable ML algorithm to predict hypoxemia in surgical patients in real time [71]. Similarly, variables associated with surgeon's performance such as distraction and cognitive load with possible implications on patient safety can be monitored using electroencephalography (EEG) [72].

Minimally invasive surgery (MIS) lends itself very well to computer vision alerting the surgeon about danger zones and minimizing errors. AI can also aid in better 3D visualization, environment mapping, and augmented reality based visual guidance during MIS [73]. There has been a rapid shift toward adoption of robotic surgery from just 2% in 2012 to 15% in 2018 for common surgical procedures [74]. This has also spurred interest in automation of surgery. Shademan et al. demonstrated superiority of a robot (Smart Tissue Autonomous Robot) compared to surgeons for *in vivo* porcine intestinal anastomosis [75]. These advances may eventually lead to supervised autonomy for robots for removing tumors especially in routine and uncomplicated cases. AI has also shown promise in the post-operative setting by outperforming clinician judgment in triaging patients for intensive care. Also, prediction of postoperative morbidity like infections has been demonstrated [76, 77]. Surgical care is a major component of cancer care and will inevitably be a beneficiary of AI-related advances in surgery.

- **Role in Radiation Oncology:** Due to heavy reliance on digital data and computer software, radiation oncology is particularly ripe for disruption by AI-based technologies [78]. Clinical decision making can be influenced by AI systems estimating the likelihood of benefit from radiation [79, 80]. After the decision to radiate is made, next step is planning and preparation. One of the most crucial steps is image

segmentation that is performed manually by the radiation oncologist. There is significant interobserver bias in this process and may result in over or under radiation [81]. AI algorithms have been developed that perform this segmentation task with accuracy rivaling experts in select cancers [82, 83]. Successful AI-based delineation of adjacent sensitive structures especially in the head and neck, mediastinal, and abdominal region has been demonstrated [84, 85]. AI algorithms have also shown promise in accelerating the process of dose calculation and optimal radiation distribution [86, 87]. During the treatment itself, application of AI to minimize effect of motion and adapt to changes in anatomy have been studied [88, 89]. Finally, the effectiveness of treatment delivered and complication prediction including mucositis, dysphagia, and rectal toxicity have also been examined with promising results [90–92]. Despite the potential benefits, external validation of the proof of concept studies remains the biggest challenge for AI in radiation oncology.

Conclusion

Overall, AI is poised to disrupt the entire cancer care cycle. Much of the progress will depend on developing innovative frameworks that enable optimization of care delivery [35]. With better – and diverse – training datasets, improving ML techniques, and greater intellectual investment from the academic oncology community, it is entirely possible to enhance the current state – and capacity – of cancer care. Technological innovation in the artificial intelligence developers' community will not stop – the question is how willing the medical community is on integrating these innovations into routine cancer care.

References

- Turing AM. Computing machinery and intelligence. Parsing the Turing test. Springer; 2009. p. 23–65.
- Smith C, McGuire B, Huang T, Yang G. The history of artificial intelligence. University of Washington; 2006. p. 27.
- Miaile N. Understanding the rise of artificial intelligence. Introduction. *Field Actions Sci Rep J Field Actions*. 2017;(Special Issue 17):5. <https://journals.openedition.org/factsreports/4382>.
- Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol*. 2017;2(4):230–43.
- Yang W, Williams JH, Hogan PF, Bruinooge SS, Rodriguez GI, Kosty MP, et al. Projected supply of and demand for oncologists and radiation oncologists through 2025: an aging, better-insured population will result in shortage. *J Oncol Pract*. 2014;10(1):39–45.
- Chen J, Remulla D, Nguyen JH, Aastha D, Liu Y, Dasgupta P, et al. Current status of artificial intelligence applications in urology and their potential to influence clinical practice. *BJU Int*. 2019;124(4):567–77.
- Nyce C, Cpcu A. Predictive analytics white paper. American Institute for CPCU. Insurance Institute of America. 2007:9–10.
- Szeliski R. Computer vision: algorithms and applications. Springer Science & Business Media; 2010.
- Cambray E, White B. Jumping NLP curves: a review of natural language processing research. *IEEE Comput Intell Mag*. 2014;9(2):48–57.
- Cook BL, Progovac AM, Chen P, Mullin B, Hou S, Baca-Garcia E. Novel use of natural language processing (NLP) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in Madrid. *Comput Math Methods Med*. 2016;2016:8708434.
- Gormley EA, Lightner DJ, Faraday M, Vasavada SP. Diagnosis and treatment of overactive bladder (non-neurogenic) in adults: AUA/SUFU guideline amendment. *J Urol*. 2015;193(5):1572–80.
- Alpaydin E. Introduction to machine learning. MIT Press; 2020.
- Aggarwal CC. Neural networks and deep learning. Springer; 2018.
- Avanzo M, Stancanello J, El Naqa I. Beyond imaging: the promise of radiomics. *Phys Med*. 2017;38: 122–39.
- Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: the process and the challenges. *Magn Reson Imaging*. 2012;30(9): 1234–48.
- Liu Z, Wang S, Di Dong JW, Fang C, Zhou X, Sun K, et al. The applications of radiomics in precision diagnosis and treatment of oncology: opportunities and challenges. *Theranostics*. 2019;9(5):1303.
- Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. *Acta Oncol*. 2010;49(7):1012–6.
- Ger RB, Zhou S, Chi P-CM, Lee HJ, Layman RR, Jones AK, et al. Comprehensive investigation on controlling for CT imaging variabilities in radiomics studies. *Sci Rep*. 2018;8(1):1–14.
- Shafiq-ul-Hassan M, Zhang GG, Latifi K, Ullah G, Hunt DC, Balagurunathan Y, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med Phys*. 2017;44(3):1050–62.

20. Yang F, Dogan N, Stoyanova R, Ford JC. Evaluation of radiomic texture feature error due to MRI acquisition and reconstruction: a simulation study utilizing ground truth. *Phys Med.* 2018;50:26–36.
21. Larue RT, Defraene G, De Ruysscher D, Lambin P, Van Elmpt W. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. *Br J Radiol.* 2017;90(1070):20160665.
22. Armato SG III, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med Phys.* 2011;38(2):915–31.
23. Nabi J. How bioethics can shape artificial intelligence and machine learning. *Hast Cent Rep.* 2018;48(5):10–3.
24. Lambin P, Leijenaar RT, Deist TM, Peerlings J, De Jong EE, Van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol.* 2017;14(12):749–62.
25. Gomolin A, Netchiporuk E, Gniadecki R, Litvinov IV. Artificial intelligence applications in dermatology: where do we stand? *Front Med.* 2020;7:100. <https://doi.org/10.3389/fmed.2020.00100>
26. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 2017;542(7639):115–8.
27. Xie F, Fan H, Li Y, Jiang Z, Meng R, Bovik A. Melanoma classification on dermoscopy images using a neural network ensemble model. *IEEE Trans Med Imaging.* 2016;36(3):849–58.
28. Yu L, Chen H, Dou Q, Qin J, Heng P-A. Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Trans Med Imaging.* 2016;36(4):994–1004.
29. Chang W-Y, Huang A, Yang C-Y, Lee C-H, Chen Y-C, Wu T-Y, et al. Computer-aided diagnosis of skin lesions using conventional digital photography: a reliability and feasibility study. *PLoS One.* 2013;8(11):e76212.
30. Spyridonos P, Gaitanis G, Likas A, Bassukas ID. Automatic discrimination of actinic keratoses from clinical photographs. *Comput Biol Med.* 2017;88:50–9.
31. Marchetti MA, Codella NC, Dusza SW, Gutman DA, Helba B, Kalloo A, et al. Results of the 2016 international skin imaging collaboration international symposium on biomedical imaging challenge: comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. *J Am Acad Dermatol.* 2018;78(2):270–7.e1.
32. Marchetti MA, Liopyris K, Dusza SW, Codella NC, Gutman DA, Helba B, et al. Computer algorithms show potential for improving dermatologists' accuracy to diagnose cutaneous melanoma: results of the international skin imaging collaboration 2017. *J Am Acad Dermatol.* 2020;82(3):622–7.
33. Dhane DM, Maity M, Mungle T, Bar C, Achar A, Kolekar M, et al. Fuzzy spectral clustering for automated delineation of chronic wound region using digital images. *Comput Biol Med.* 2017;89:551–60.
34. García-Zapirain B, Elmogy M, El-Baz A, Elmaghhraby AS. Classification of pressure ulcer tissues with 3D convolutional neural network. *Med Biol Eng Comput.* 2018;56(12):2245–58.
35. Nabi J. Addressing the “wicked” problems in machine learning applications – time for bioethical agility. *Am J Bioeth.* 2020;20(11):25–7.
36. Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *J Investig Dermatol.* 2018;138(7):1529–38.
37. Moxley-Wyles B, Colling R, Verrill C. Artificial intelligence in pathology: an overview. *Diagn Histopathol.* 2020;26:513.
38. Zarella MD, Bowman D, Aeffner F, Farahani N, Xthona A, Absar SF, et al. A practical guide to whole slide imaging: a white paper from the digital pathology association. *Arch Pathol Lab Med.* 2019;143(2):222–34.
39. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology – new tools for diagnosis and precision oncology. *Nat Rev Clin Oncol.* 2019;16(11):703–15.
40. Liu Y, Gadepalli K, Norouzi M, Dahl GE, Kohlberger T, Boyko A, et al. Detecting cancer metastases on gigapixel pathology images. *arXiv preprint arXiv:170302442.* 2017.
41. Ozkan TA, Eruyar AT, Cebeci OO, Memik O, Ozcan L, Kuskomuz I. Interobserver variability in Gleason histological grading of prostate cancer. *Scand J Urol.* 2016;50(6):420–4.
42. Nagpal K, Foote D, Liu Y, Chen P-HC, Wulczyn E, Tan F, et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ Digit Med.* 2019;2(1):1–10.
43. Kim RH, Nomikou S, Dawood Z, Jour G, Donnelly D, Moran U, et al. A deep learning approach for rapid mutational screening in melanoma. *bioRxiv.* 2019;610311. <https://doi.org/10.1101/610311>
44. Schaumberg AJ, Rubin MA, Fuchs TJ. H&E-stained whole slide image deep learning predicts SPOP mutation state in prostate cancer. *BioRxiv.* 2018;064279. <https://doi.org/10.1101/064279>
45. Bychkov D, Linder N, Turkki R, Nordling S, Kovanen PE, Verrill C, et al. Deep learning based tissue analysis predicts outcome in colorectal cancer. *Sci Rep.* 2018;8(1):1–11.
46. Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci Rep.* 2016;6:26286.
47. Shimizu H, Nakayama KI. Artificial intelligence in oncology. *Cancer Sci.* 2020;111(5):1452.
48. Ahmed Z, Mohamed K, Zeeshan S, Dong X. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Databse.* 2020;2020:baaa010.
49. Wang Y-H, Nguyen PA, Islam MM, Li Y-C, Yang H-C, editors. *Development of deep learning algorithm for*

- detection of colorectal cancer in EHR data. MedInfo; 2019.
50. Shao C-H, Chen C-L, Lin J-Y, Chen C-J, Fu S-H, Chen Y-T, et al. Metabolite marker discovery for the detection of bladder cancer by comparative metabolomics. *Oncotarget*. 2017;8(24):38802.
 51. Zheng H, Ji J, Zhao L, Chen M, Shi A, Pan L, et al. Prediction and diagnosis of renal cell carcinoma using nuclear magnetic resonance-based serum metabolomics and self-organizing maps. *Oncotarget*. 2016;7(37):59189.
 52. Gerstung M, Jolly C, Leshchiner I, Dentro SC, Gonzalez S, Rosebrock D, et al. The evolutionary history of 2,658 cancers. *Nature*. 2020;578(7793): 122–8.
 53. Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat Genet*. 2018;50(8):1171–9.
 54. Schilsky RL, Nass S, Le Beau MM, Benz EJ Jr. Progress in cancer research, prevention, and care. *N Engl J Med*. 2020;383(10):897–900.
 55. Vanneman M, Dranoff G. Combining immunotherapy and targeted therapies in cancer treatment. *Nat Rev Cancer*. 2012;12(4):237–51.
 56. Liang G, Fan W, Luo H, Zhu X. The emerging roles of artificial intelligence in cancer drug development and precision therapy. *Biomed Pharmacother*. 2020;128: 110255.
 57. Xu J, Yang P, Xue S, Sharma B, Sanchez-Martin M, Wang F, et al. Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives. *Hum Genet*. 2019;138(2):109–24.
 58. Somashekhar S, Sepúlveda M-J, Puglielli S, Norden A, Shortliffe E, Rohit Kumar C, et al. Watson for oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. *Ann Oncol*. 2018;29(2):418–23.
 59. Somashekhar S, Sepúlveda M-J, Shortliffe EH, Rauthan A, Patil P, Yethadka R. A prospective blinded study of 1000 cases analyzing the role of artificial intelligence: Watson for oncology and change in decision making of a multidisciplinary tumor board (MDT) from a tertiary care cancer center. American Society of Clinical Oncology; 2019.
 60. Reyes-Aldasoro CC. The proportion of cancer-related entries in PubMed has increased considerably: is cancer truly “the emperor of all maladies”? *PLoS One*. 2017;12(3):e0173671.
 61. Williams AM, Liu Y, Regner KR, Jotterand F, Liu P, Liang M. Artificial intelligence, physiological genomics, and precision medicine. *Physiol Genomics*. 2018;50(4):237–43.
 62. Patel NM, Michelini VV, Snell JM, Balu S, Hoyle AP, Parker JS, et al. Enhancing next-generation sequencing-guided cancer care through cognitive computing. *Oncologist*. 2018;23(2):179.
 63. Forbes SA, Beare D, Boutsikakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res*. 2017;45(D1):D777–D83.
 64. Wang Y, Wang Z, Xu J, Li J, Li S, Zhang M, et al. Systematic identification of non-coding pharmacogenomic landscape in cancer. *Nat Commun*. 2018;9(1):1–15.
 65. Pantuck AJ, Lee DK, Kee T, Wang P, Lakhotia S, Silverman MH, et al. Modulating BET bromodomain inhibitor ZEN-3694 and enzalutamide combination dosing in a metastatic prostate cancer patient using CURATE. AI, an artificial intelligence platform. *Adv Ther*. 2018;1(6):1800104.
 66. Sun R, Limkin EJ, Vakalopoulou M, Dercle L, Champiat S, Han SR, et al. A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study. *Lancet Oncol*. 2018;19(9):1180–91.
 67. Bahl M, Barzilay R, Yedidia AB, Locascio NJ, Yu L, Lehman CD. High-risk breast lesions: a machine learning model to predict pathologic upgrade and reduce unnecessary surgical excision. *Radiology*. 2018;286(3):810–8.
 68. Hu L, Bell D, Antani S, Xue Z, Yu K, Horning MP, et al. An observational study of deep learning and automated evaluation of cervical images for cancer screening. *J Natl Cancer Inst*. 2019;111(9):923–32.
 69. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial intelligence in surgery: promises and perils. *Ann Surg*. 2018;268(1):70.
 70. Xia J, Samman N, Yeung RW, Wang D, Shen SG, Ip HH, et al. Computer-assisted three-dimensional surgical planning and simulation: 3D soft tissue planning and prediction. *Int J Oral Maxillofac Surg*. 2000;29(4): 250–8.
 71. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng*. 2018;2(10):749–60.
 72. Guru KA, Shafiei SB, Khan A, Hussein AA, Sharif M, Esfahani ET. Understanding cognitive performance during robot-assisted surgery. *Urology*. 2015;86(4): 751–7.
 73. Zhou X-Y, Guo Y, Shen M, Yang G-Z. Application of artificial intelligence in surgery. *Front Med*. 2020;14: 417–30.
 74. Sheetz KH, Claffin J, Dimick JB. Trends in the adoption of robotic surgery for common surgical procedures. *JAMA Netw Open*. 2020;3(1):e1918911-e.
 75. Shademan A, Decker RS, Opfermann JD, Leonard S, Krieger A, Kim PC. Supervised autonomous robotic soft tissue surgery. *Sci Transl Med*. 2016;8(337): 337ra64-ra64.
 76. Hopkins BS, Mazmudar A, Driscoll C, Svet M, Goergen J, Kelsten M, et al. Using artificial intelligence (AI) to predict postoperative surgical site infection: a

- retrospective cohort of 4046 posterior spinal fusions. *Clin Neurol Neurosurg.* 2020;192:105718.
77. Hsieh N-C, Hung L-P, Shih C-C, Keh H-C, Chan C-H. Intelligent postoperative morbidity prediction of heart disease using artificial intelligence techniques. *J Med Syst.* 2012;36(3):1809–20.
78. Huynh E, Hosny A, Guthier C, Bitterman DS, Petit SF, Haas-Kogan DA, et al. Artificial intelligence in radiation oncology. *Nat Rev Clin Oncol.* 2020;17(12):771–81.
79. Kann BH, Aneja S, Loganadane GV, Kelly JR, Smith SM, Decker RH, et al. Pretreatment identification of head and neck cancer nodal metastasis and extranodal extension using deep learning neural networks. *Sci Rep.* 2018;8(1):1–11.
80. Deist TM, Dankers FJ, Valdes G, Wijsman R, Hsu IC, Oberije C, et al. Machine learning algorithms for outcome prediction in (chemo) radiotherapy: an empirical comparison of classifiers. *Med Phys.* 2018;45(7):3449–59.
81. Ohri N, Shen X, Dicker AP, Doyle LA, Harrison AS, Showalter TN. Radiotherapy protocol deviations and clinical outcomes: a meta-analysis of cooperative group clinical trials. *J Natl Cancer Inst.* 2013;105(6):387–93.
82. Cardenas CE, McCarroll RE, Court LE, Elghohari BA, Elhalawani H, Fuller CD, et al. Deep learning algorithm for auto-delineation of high-risk oropharyngeal clinical target volumes with built-in dice similarity coefficient parameter optimization function. *Int J Radiat Oncol Biol Phys.* 2018;101(2):468–78.
83. Men K, Chen X, Zhang Y, Zhang T, Dai J, Yi J, et al. Deep deconvolutional neural network for target segmentation of nasopharyngeal cancer in planning computed tomography images. *Front Oncol.* 2017;7:315.
84. Jackson P, Hardcastle N, Dawe N, Kron T, Hofman MS, Hicks RJ. Deep learning renal segmentation for fully automated radiation dose estimation in unsealed source therapy. *Front Oncol.* 2018;8:215.
85. Lustberg T, van Soest J, Gooding M, Peressutti D, Aljabar P, van der Stoep J, et al. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. *Radiother Oncol.* 2018;126(2):312–7.
86. Nguyen D, Long T, Jia X, Lu W, Gu X, Iqbal Z, et al. A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning. *Sci Rep.* 2019;9(1):1–10.
87. Xing Y, Nguyen D, Lu W, Yang M, Jiang S. A feasibility study on deep learning-based radiotherapy dose calculation. *Med Phys.* 2020;47(2):753–8.
88. Guidi G, Maffei N, Meduri B, D'Angelo E, Mistretta G, Ceroni P, et al. A machine learning tool for re-planning and adaptive RT: a multicenter cohort investigation. *Phys Med.* 2016;32(12):1659–66.
89. Isaksson M, Jalden J, Murphy MJ. On using an adaptive neural network to predict lung tumor motion during respiration for radiotherapy applications. *Med Phys.* 2005;32(12):3801–9.
90. Dean J, Wong K, Gay H, Welsh L, Jones A-B, Schick U, et al. Incorporating spatial dose metrics in machine learning-based normal tissue complication probability (NTCP) models of severe acute dysphagia resulting from head and neck radiotherapy. *Clin Transl Radiat Oncol.* 2018;8:27–39.
91. Dean JA, Wong KH, Welsh LC, Jones A-B, Schick U, Newbold KL, et al. Normal tissue complication probability (NTCP) modelling using spatial dose metrics and machine learning methods for severe acute oral mucositis resulting from head and neck radiotherapy. *Radiother Oncol.* 2016;120(1):21–7.
92. Gopalakrishnan K, Khaitan SK, Choudhary A, Agrawal A. Deep convolutional neural networks with transfer learning for computer vision-based data-driven pavement distress detection. *Constr Build Mater.* 2017;157:322–30.



Artificial Intelligence in Radiotherapy and Patient Care

91

James Chun Lam Chow

Contents

Introduction	1276
Basic Concept of AI and Machine Learning	1277
Radiotherapy Chain	1277
Applications of AI and Machine Learning in Radiotherapy	1278
Radiation Treatment Planning	1278
Treatment Plan Evaluation	1280
Radiation Dose Delivery Using Multileaf Collimator	1282
Chatbot	1283
Conclusion and Future Prospective	1285
References	1285

Abstract

Artificial intelligence (AI) is the intelligence of a machine to perform a task. To date, with the rapid advances of machine learning algorithm, big data analytics, cloud computing, and mobile network, AI has been studied and used in radiotherapy and patient care. AI and machine learning contribute to the automation,

process enhancement, decision-making, effective resource allocation, and medical education in different aspects of radiotherapy. These, not least, include automations in the radiation treatment planning regarding auto-contouring, auto-beam setup, and prediction of dose distribution; plan evaluation and quality assurance (QA) using big data cloud and machine learning; radiation dose delivery using AI-assisted leaf-sequencing for the multileaf collimator; and chatbot for patient care and radiotherapy education. In this chapter, we review the background and basic concept of AI and machine learning, and how to implement this digital tool to automate the radiotherapy chain. We explore some recent ideas and results on the applications of AI and machine learning such as treatment planning and QA, radiation dose

J. C. L. Chow (✉)

Radiation Medicine Program, Princess Margaret Cancer Centre, University Health Network, Toronto, ON, Canada
Department of Radiation Oncology, University of Toronto, Toronto, ON, Canada

TECHNA Institute for the Advancement of Technology for Health, University Health Network, Toronto, ON, Canada
e-mail: james.chow@rmpuhn.ca

© Springer Nature Switzerland AG 2022

N. Lidströmer, H. Ashrafiyan (eds.), *Artificial Intelligence in Medicine*,
https://doi.org/10.1007/978-3-030-64573-1_143

1275

delivery, and chatbot. We also discuss potential challenges related to the implementation of AI in radiotherapy and its future prospective.

Keywords

Artificial intelligence · Machine learning · Radiotherapy · Treatment planning · Chatbot · Healthcare · Quality assurance · Big data · Cloud computing · High-performance computing

Introduction

Artificial intelligence (AI) is an attractive topic. Just imagine a machine having human intelligence to complete a task. With rapid computer advances of high-performance computing, big data analytics, and machine learning algorithms in medicine [1–3], applying this digital tool in radiation treatment [4] is the way to go. Radiotherapy is a type of cancer treatment using ionizing radiations such as photon, electron, and proton beams. The therapy consists of a chain of processes including disease diagnosis and dose prescription, treatment simulation, treatment planning, plan evaluation and quality assurance (QA), beam delivery, and patient follow-up. Since most processes involve computer medical image processing, calculation, and hardware control, AI and machine learning are used in the radiotherapy chain. To date, some AI and machine learning systems have been proposed, studied, and developed to automate the treatment process, enhance the resource allocation, and increase the accuracy and speed of decision-making in radiotherapy [4–6]. The AI system facilitates the radiotherapy process, making the radiation dose delivery more accurate and precise. In this chapter, we will explore some recent results in the application of AI and machine learning in radiotherapy, focusing particularly on the treatment planning, plan evaluation and QA, dose delivery, and patient education and care.

The aim of radiotherapy is to reduce the size of tumor and kill the cancer cell. This can be helped by a radiation treatment plan aiming at maximizing the tumor target dose coverage, while sparing the normal tissues or organs-at-risk. Therefore,

highlighting or contouring the target (tumor) and organs-at-risk from the patient's anatomy three-dimensionally based on the medical image set is important. In this chapter, we will review the progress of using AI to assist auto-contouring of target and critical organs in treatment planning. Results of how machine learning can improve the accuracy and speed of contouring in the treatment planning system will be discussed. For the calculation of dose in the irradiated volume of patient in a treatment plan, AI can help to predict the dose distribution according to the previous results, provided that the radiation beam geometry and treatment site are similar. Through machine training, accurate dose distribution for a treatment plan can be predicted quickly without performing dose calculation.

Taking advantage of high-performance and parallel computing, hundreds of treatment plans can be created within minutes using a treatment planning system. These plans contain different dose-volume information of the target and organs-at-risk, radiobiological variables, beam parameters such as type, energy, field size, and angle, and delivery techniques such as intensity modulated radiotherapy and volumetric modulated arc therapy. Justifying the best plan for the patient in the treatment requires expertise and experience of the treatment team of radiation oncologist, medical physicist, and radiotherapist. Using AI and machine learning, the dose distribution index (DDI) showing important dose distribution information in the treatment plan can be predicted rather than calculated [7, 8]. From the predicted DDI, the quality of a treatment plan regarding the target coverage and organs-at-risk sparing can be evaluated. We will also explore a comprehensive AI-assisted QA system supported by Internet-of-things (IoT) [9].

Intensity modulated radiotherapy is a popular and advanced radiation dose delivery method using multileaf collimator attached to the head of a medical linear accelerator. The multileaf collimator contains individual “leaves” which are able to move independently in and out of the path of the radiation beam. The intensity of the beam can therefore be modulated by the beam segments generated by the collimator. Generally, the dose coverage of the target can be delivered by a group of beam segments at different angles or around an

arc [10]. The shape of the beam segment can be calculated by a leaf-sequencing algorithm based on a mathematical model. Using AI and machine learning, however, related beam segments can be determined without using traditional mathematical model. This makes the dose delivery process less complicated, faster, and more efficient.

For senior cancer patients who need extra care and support, maintaining good communication is a must for knowledge transfer regarding patient education and care. With the limitation of human resource, a humanlike AI-assisted chatbot would be ideal in this healthcare issue [11]. Chatbot running with natural language processing can understand the need from the cancer patients and direct them to the right path to obtain help [12]. Chatbot can also “talk” to the patient for education in radiotherapy and information acquisition such as doing a survey. Using chatbot for patient communication in radiotherapy can highly reduce resource and increase patient satisfaction. This is because the chatbot is working untiringly 24/7 with user-friendly humanlike manner on all IoT such as tablet, smartphone, and computer.

We will review the basic principle of AI and machine learning, and understand every process in the radiotherapy chain. We will explore recent applications of AI in some processes in the radiotherapy chain. Finally, we will conclude the present progress and foresee the future prospective of AI and machine learning in radiotherapy.

Basic Concept of AI and Machine Learning

AI is generally regarded as a computer system which performs functions that are normally performed by humans and is used to describe a computer or machine which can mimic human learning and problem solving [13]. The modern idea of AI was initialized by people attempting to investigate the process of human thinking as the mechanical manipulation of symbols. Still, due to various understandings of concepts and terms of AI in different fields such as business, science, engineering, music, medicine, and so on, there are different definitions and descriptions of AI. These mainly refer to software systems that provide

decision-making or think autonomously like human experts. For application of AI in medicine that clinicians aim at providing the best possible medical care to patients, AI is expected to improve clinical outcomes and reduce resource [14]. Any developments of computer hardware and software regarding AI that can help clinicians to achieve the above goals will be desired.

Machine learning is a computer algorithm which do not require traditional programming with rules to perform their functions. Machine learning is a subset of AI and its algorithms can perform a task through training themselves by analyzing a big dataset [15]. An example of machine learning in medicine is to assist the dermatologists in diagnosing melanoma [16]. The computer system is trained to distinguish between skin cancer and a normal mole after analyzing more than 100,000 patient images. Although machine learning takes many thousands more of viewing to recognize an object, while human only needs a couple of viewing, machine learning algorithms after training can detect subtle differences in millions of pixels in a dermatology image set, which is almost invisible to the human eye. Neural network are algorithms designed to mimic the functioning of the human nervous systems [17]. In human, the neural network contains neurons and synapses, while in a computer system, the network is connected by artificial neurons or nodes. To simulate the operation of human brain managing a large amount of data such as digital image, each node in the computer neural network is excited by the data from the image, then sent to the next node. The excitement transferred from one node to another is represented by a number or weight in the algorithm. The goal is to transfer the signal from the input to output through different layers of nodes as shown in Fig. 1.

Radiotherapy Chain

Radiotherapy is a cancer treatment using ionizing radiation to control or kill cancer cells. The whole treatment process can be described by the radiotherapy chain as shown in Fig. 2.

It is seen in Fig. 2 that the radiotherapy chain includes seven steps, namely, diagnosis and

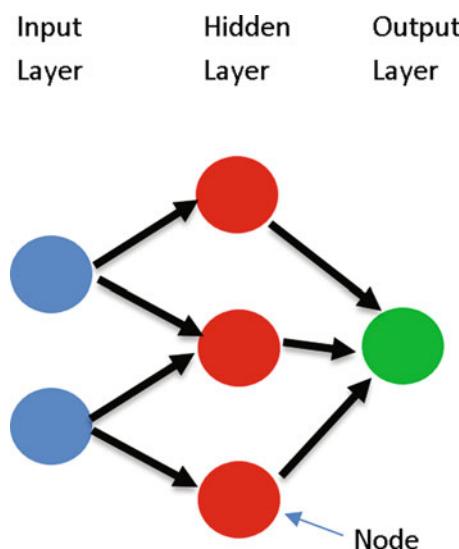


Fig. 1 Schematic diagram of a simple neural network

prescription, patient positioning, immobilization and simulation, target and organ contouring, radiation treatment planning, setup verification, and treatment delivery and completion of treatment or patient follow-up. When a patient is sent to the cancer center, various clinical tests are carried out for diagnosis of the disease. These tests include medical imaging and laboratory tests, helping the radiation oncologist to find out the stage and grade of the tumor, and whether the cancer has spread to other organs. The radiation oncologist then prescribes the radiation dose to the target (tumor), and the patient is sent to the simulation unit for treatment simulation and planning. The treatment position of the patient is set up by the radiotherapists, and an immobilization device may be used to increase the positional accuracy. A computed tomography (CT) scan is performed by a CT-simulator to acquire the 3D-anatomy information of the patient. From the CT image set, the radiation oncologist contours the target and organs-at-risk in the treatment planning system. The contoured image set is followed by a planner or medical dosimetrist to create a treatment plan using radiation beams with various type, field size, energy, and angle. Some advanced dose delivery techniques such as intensity-modulated radiotherapy or volumetric modulated radiotherapy can be used [18]. The created treatment plan is

reviewed and approved by the radiotherapist, medical physicist, and radiation oncologist, and a patient-specific quality control test is conducted by measurement. When all treatment QA procedures are done and passed, the patient is sent to the treatment unit for dose delivery. In the course of treatment, the delivery record of the patient is reviewed by a radiotherapist weekly. Patient follow-up is started after the treatment is complete to investigate the patient outcome monthly and yearly.

Applications of AI and Machine Learning in Radiotherapy

Since most processes in the radiotherapy chain involve computer control, analysis, and calculation [1], with the recent progress made in machine learning algorithm, big data processing, and high-performance computing, AI has been studied and implemented in the chain so as to automate and enhance the efficiency of the treatment [4]. In this section, we will highlight some applications of AI in various processes in the radiotherapy chain such as treatment planning, plan evaluation and QA, dose delivery, and patient care.

Radiation Treatment Planning

In radiotherapy, radiation treatment planning is the process in which a team containing radiation oncologist, medical physicist, and radiotherapist (or medical dosimetrist) plan the appropriate external beam radiotherapy or internal brachytherapy treatment technique for a patient with cancer. Implementing AI in radiation treatment planning is a long-term project requiring design of computer algorithm and infrastructure of application, set up of corresponding clinical policies, purchase and development of software and hardware, and national and international collaboration. Although such implementation process is time-consuming and complicated, some recent results are still noticeable.

In treatment planning, generating a 3D-dose distribution of a plan requires a time-consuming

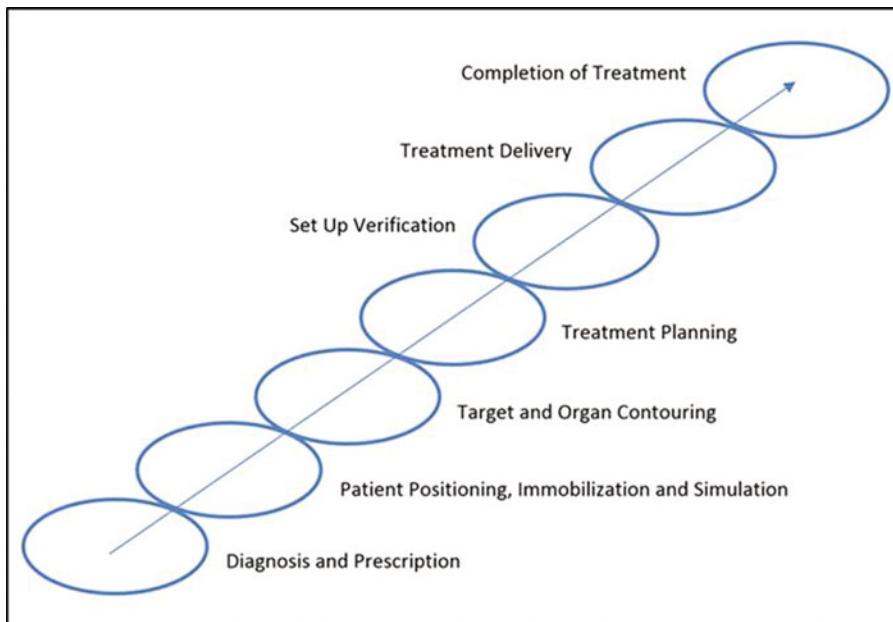


Fig. 2 Radiotherapy chain showing different steps in the radiation treatment process

calculation using various methods such as Monte Carlo simulation or convolution-superposition algorithm [19, 20]. With machine learning, the dose distribution can be determined through training associated with a pipeline of voxel-based dose prediction algorithm [21]. After machine training with previously treated patients based on multi-patient atlas selection, the system can learn to automatically select the most effective atlases for a new patient. The planning system can match the dose from those atlases onto the new patient. This AI process can avoid the time-consuming dose calculation step, and it is particularly useful when dealing with a huge number of patients having the same treatment site (e.g., breast, CNS brain, prostate, lung, and rectum), using the same delivery technique (e.g., intensity-modulated radiotherapy and volumetric modulated arc therapy) and treatment protocol.

Auto-contouring or image segmentation can be assisted by machine learning to increase the accuracy and efficiency in treatment planning. Atlas-based auto-segmentation is built and implemented to the treatment planning system [22]. This auto-segmentation is an AI-assisted process to perform segmentation on novel data using a dataset, which

consists of structures-of-interest from previous segmentation. The registration efficiency and robustness of the segmentation are improved by the registration strategy combined with the objects' shape information in the atlas. By comparing results of manual contour, atlas-based contour, and user-adjusted deep learning contour, it is found that the deep learning contour method performed the best on the lung, esophagus, spinal cord, heart, and mediastinum contours [23].

Machine learning can also help in beam angle optimization in treatment planning. In lung intensity-modulated radiotherapy, an automatic beam selection method is developed using machine learning [24]. A beam selection score function is built by considering the anatomical and beam features of the radiation beams. These include the planning target volume distribution in the thorax, the planning target volume projection shape, beam distances, and projected target volume and organs-at-risk overlap. The automatic beam selection method was tested using 149 plans with 122 for training and 26 for verifying [25]. Three to eight radiation beams were used in the plan. Results showed that the clinical acceptability criteria (e.g., planning target volume

coverage, maximal spinal canal dose, and mean lung dose) of the automatic beams are comparable to the clinical beams used as a control.

Treatment Plan Evaluation

Using the new automatic treatment planning system incorporated with high-performance computing, a large number of treatment plans can be generated at the same time that satisfied the dose-volume criteria based on the treatment protocol [1, 4]. Plan evaluation is therefore selecting the best plan for the patient. This evaluation used to be done by comparing different parameters (e.g., beam delivery, radiobiological, dose distribution, and delivery technique) and requires experts (e.g., radiation oncologist and medical physicist) to justify those parameters among plans. Machine learning simplified the plan evaluation process using a big data cloud [26]. In the following sections, we will explore an AI-assisted QA system using IoT for plan evaluation, followed by a computer system using machine learning to predict DDI from dose-volume histograms (DVHs) [27].

Treatment Plan QA

QA is a program designed to control and maintain a standard of quality to avoid and eliminate errors. This program is important in every step of the radiotherapy chain especially treatment planning, with recent implementation of large-scale clinical dose reconstruction and adaptive radiotherapy. A novel automatic QA system using machine learning for IoT is necessary to create [1, 28]. IoT refers to a network of physical devices which can interchange data. In radiotherapy, IoT includes the medical-imaging modules, treatment-planning system, image-guided system, and medical linear accelerator console. For treatment plan QA, data is exchanged among IoT to generate updated plan parameters for evaluation. This includes dose reconstruction results predicted using cone-beam CT [29]. Radiation staff are required to carry out real-time QA before every daily dose delivery in the treatment unit. This brings up two problems: (1) Many radiation experts are needed to perform

personalized QA; and (2) despite the treatment QA guidelines, evaluation results often differ among different radiation staff. To solve these problems, AI will be employed in the QA program in the radiotherapy chain with nonbiased justification.

To apply AI on IoT to automate the QA program in the radiotherapy chain, an IoT-enabled radiotherapy chain can be formed based on all clinical devices connected to the hospital network. In an IoT-enabled radiotherapy chain, the image set acquired from the CT-simulator is sent to the treatment planning system through intranet. There are five steps in the radiotherapy process: (1) A treatment plan is created with specific beam energy, geometry, and delivery technique based on the image set from the CT-simulator; (2) radiation dose is calculated based on the plan data, producing dosimetric results in form of DVHs, dose-volume points, and radiobiological parameters; (3) plan evaluation is carried out by the radiation oncologist and medical physicist; (4) when the plan is approved, treatment planning data is sent through the data network to the dosimetric QA systems for personalized treatment QA; and (5) when the QA is finished and approved, the plan data is sent to the medical linear accelerator console for radiation dose delivery. During these five steps, there is a lot of data transferred in the IoT chain involving several complicated systems, for example, from the CT-simulator to the treatment planning system and finally to the linear accelerator. Without setting them up in an IoT chain, it takes a very long time to go manually from the start to the end of the radiotherapy system.

AI plus IoT in radiotherapy can solve a huge bottleneck in the cone-beam CT. When a patient undergoes cone-beam CT, radiotherapists do a new quick calculation basing on the current physical dimensions of the patient, as the calculations done when the treatment first started may have changed, because the patient may become fatter or thinner. Thus a quick calculation at the cone-beam CT has to do with the patient lying on the couch in the treatment unit. This calculation necessarily has to be very quick with the patient's presence. All these mean, in image-guided adaptive radiotherapy, an image set of a patient is acquired by a

cone-beam CT system in the treatment unit before beam on. Then this most up-to-date anatomy information is compared to the planned CT image set in treatment planning to avoid any unacceptable deviation in the patient setup. If the calculations show deviations, the patient's physical position has to be corrected.

AI plus IoT can personalize the QA. Recent advancements in image-guided adaptive radiotherapy using cone-beam CT have made it possible for image processing/registration, high-performance computing on the dose distribution, and dose reconstruction to be carried out within 2–3 min [30]. Thus, the cone-beam CT image set done at the treatment unit can be used as the base to recalculate the dose distribution specifically for a treatment fraction. QA is therefore personalized, and we can predict the dose distribution in the patient if we proceed to finish the dose delivery. The radiation staff can then decide whether the treatment can go on.

The automation of the treatment plan QA program using AI and IoT can fit into the contemporary adaptive radiotherapy chain. With the AI-enabled IoT chain, radiation staff only need to do the approval once in a full course of treatment including multiple fractions (typically 5–30 fractions per treatment depending on the sites), instead of approving each step in the process.

Dose Distribution Index Prediction

To evaluate the quality of a treatment plan, various dose-volume and radiobiological variables are used to justify whether the created plan would satisfy the requirement of a corresponding clinical protocol [31, 32]. Among all plan evaluation variables, DDI is an effective one calculated by the DVHs from the target and organs-at-risk. This index not only provides the dosimetric estimates on the target coverage, but also spares all organs-at-risk and the remaining healthy tissue in the treated organ in a single variable [7]. DDI has been proved effective in treatment planning QA [8]. Here, instead of calculating the DDI using mathematical formulas based on the definition, machine learning is used to predict the DDI value from the DVHs generated by the treatment planning system [33].

To investigate machine learning in predicting the DDI for treatment plan evaluation, different algorithms used in machine training are investigated. These included linear regression, tree regression, support vector machine (SVM), and Gaussian process regression (GPR) [34]. Data of 50 prostate volumetric modulated arc therapy plans were used in machine training [35]. The DVHs for the targets (planning target and clinical target volumes) and organs-at-risk (rectal wall, bladder wall, and left and right femur) were pre-processed and transformed as five sampling points, namely, $D_{100\%}$, $D_{75\%}$, $D_{50\%}$, $D_{25\%}$, and D_M for the input of training. The workflow of the machine learning process (Fig. 3) included data collection, data preprocessing, model selection, training and validation, evaluation, and final justification. To compare the performance of the machine learning algorithms, root mean square error (RMSE) showing the deviation between the predicted and calculated DDI values, the prediction time of the machine learning and the training time were determined and compared.

Comparing the RMSE values among all machine learning algorithms, only the DDI predicted by the medium and coarse tree regression

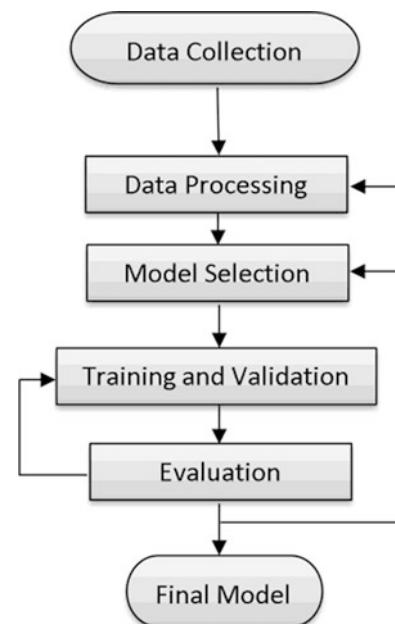


Fig. 3 Flow chart showing the machine learning workflow to determine the DDI

algorithms showed a relatively large RMSE value in the range of 0.021–0.034. These results showed that tree regression algorithms were not able to predict the DDI accurately. For other algorithms such as SVM and GRP, they all performed very well in predicting the DDI with smaller RMSE values ranging from 0.0038 to 0.0193. By considering other factors such as prediction speed and training time, the square exponential GPR algorithm had the smallest RMSE value of 0.0038, a relatively high prediction speed of 4,100 observations per second, and a short machine training time of 0.18 s [33]. This shows that the performance of the square exponential GPR algorithm is the best among others in predicting the DDI for treatment planning evaluation. It is seen that machine learning can be used to predict dose-volume variables such as DDI for treatment plan evaluation in radiotherapy.

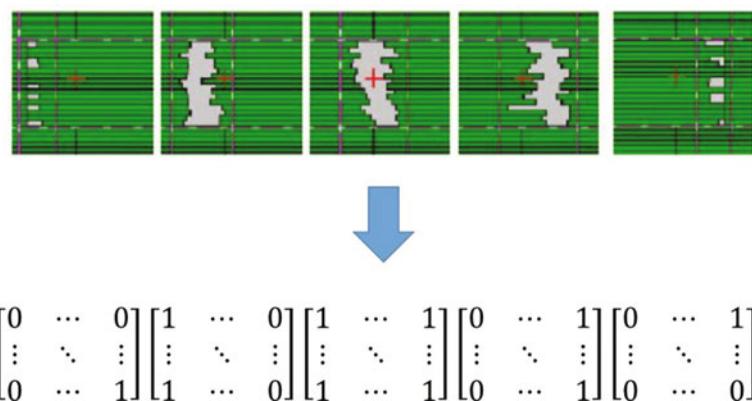
Radiation Dose Delivery Using Multileaf Collimator

In intensity-modulated radiotherapy, the tumor with irregular shape is irradiated by photon beams with various intensity-modulated profiles to achieve a dedicated dose distribution. To produce such intensity-modulated profiles, multileaf collimator is used. This computer-controlled collimator consists of a set of leaf pairs with the left and right leaves having identical size. The collimator can act as an intensity modulator to convert a given beam profile into a modulated profile or

fluence map needed in the treatment [36]. The principle is that each treatment field is divided into a group of subfields of irregular apertures (beam segments) with uniform intensity levels. These beam segments are generated by the multileaf collimator and then delivered in a stack arrangement in sequence to create the final intensity-modulated fluence map. Therefore, in intensity-modulated radiotherapy, a fluence map needs to be converted into a leaf sequence that controls the leaf movement in multileaf collimator during delivery [37]. As intensity-modulated beam techniques (static and dynamic) were developed, some leaf-sequencing algorithms were implemented in the multileaf collimator to produce beam segments of the required fluence map. These algorithms were further developed to improve the dose delivery with shorter treatment time and more accurate dosimetry [38]. Better algorithms removed more unwanted small beam segments resulting in small number of monitor unit or output produced per gantry angle in radiotherapy such as volumetric modulated arc therapy. As this small field error leading to dose distribution uncertainty can be greatly reduced, leaf-sequencing algorithm development is important in radiotherapy delivery.

Since the photon beam segment and final fluence map can be converted as graphical expressions with different intensity levels, it is possible to transform them into a grid of numbers and input them into a simple neural network. In Fig. 4, the final required fluence map was generated by five beam segments using different monitor units.

Fig. 4 Photon beam segments expressed as $m \times n$ matrices which added to become the required fluence map for intensity-modulated beam in radiotherapy



Each segment in Fig. 4 can be expressed in a $m \times n$ matrix with a constant intensity (0 or $p \times 1$), where p is related to the number of monitor units. The machine learning process is therefore carried out by feeding the beam segments in terms of graphical representations to the computer as per the targeted fluence map. The computer is trained to add the beam segments in order to produce the final fluence map required in the modulated beam dose delivery. Through machine learning, we can restrict the computer to search for the simplest combination of beam segments using the smallest number of segments, largest size of aperture, and smallest number of total monitor unit. The source of beam segments and resultant fluence map can be from the routine clinical treatment planning for the intensity-modulated radiotherapy or volumetric modulated arc therapy. There are 100,000 cases to be extracted from the treatment-planning system to a big data cloud for the machine learning process. One of the advantages in machine learning is that the leaf-sequencing process does not depend on any mathematical algorithms based on the corresponding theory and assumption. The machine learning process can find out directly the solution of the required fluence map from the beam segments with restricted condition.

Chatbot

Chatbots are the next big thing in the era of conversational services. It is a virtual person who can talk to human beings using interactive textual skills. There are currently many cloud-based chatbot development platforms such as IBM Watson, Microsoft Bot, AWS Lambda, and Heroku. In healthcare, AI-assisted chatbot is an innovative approach that can be used conveniently in health promotion and patient care interventions [39]. Conventional assistants can be deployed to IoT within different applications. Moreover, with an increased focus on the individual in health, hospital medical services have been shifting from a treatment focus to prevention and health management. The medical industry is creating additional services for health and life-promotion programs. AI-assisted chatbot is a framework

enabling a smooth human-robot interaction that supports the efficient implementation of the healthcare service. The possible ways to utilize chatbots to assist healthcare providers and support patients have been much discussed. It was concluded that chatbots will play a leading role by embodying the function of a virtual assistant and bridging the gap between patients and clinicians [40]. Moreover, AI and machine learning on chatbots can help to save healthcare costs when used in place of a human, or assist them as a preliminary step to assess a condition and provide self-care recommendation. A semistructured interview using online survey on the demographic and attitudinal variables such as acceptability and perceived utility of a chatbot was conducted [40]. Results showed that most Internet users would be receptive to using health chatbots, though hesitancy regarding this technology is likely to compromise engagement. We will focus on the recent development of chatbots for patient education and care in radiotherapy.

Background and Basic Concept of a Chatbot

When you are searching for some information about cancer treatment (e.g., what is radiotherapy?) as a member of the general public, there are many ways you can do it. You can search the information by going to the library to find some related references, look it up in some textbooks you remember placed somewhere in your study room, or calling somebody who you believe knows the answer. On the other hand, you can work otherwise using the Google search on your laptop or desktop, asking Siri on your iPhone, or posting the question to a WhatsApp group concerning the related topic and waiting for the answer. With recent rapid advances on IoT in the adoption of mobiles such as smartphone and tablet, apps are developed as specific channels for users to acquire or exchange information through the Internet. For an app providing information, the users demand instant response once they send out a question. The app should understand what the user needs immediately and provide what they are looking for accurately and precisely. This results in the innovations of chatbot powered by AI,

which is going to help the knowledge provider to automate the information transfer to the user, taking advantage of the Internet and big data.

A chatbot is a computer program designed to simulate conversation with human users through the Internet. The chatbot is an app that automates selected tasks by chatting with a user through a conversational interface. Figure 5 shows a simple workflow of the chatbot (Bot).

In Fig. 5, the user will type in a text as their enquiry or need to the chatbot interface. The chatbot will respond to the user using a question as confirmation. If the user answers “Yes.” the chatbot will process “do something” and return the information as requested by the user. If the user answers “No,” the chatbot will also carry out a task such as asking another question to further test and verify the user’s need. However, if the chatbot does not know what the user input to the interface, it will provide a guidance with some options to select in order to help the user to acquire their need.

Chatbot for Radiotherapy Education and Patient Care

For cancer patients, more than half of them receive radiotherapy as part of their treatment. The radiotherapy process is complex and complicated, as it is multidisciplinary involving knowledge in radiation oncology, medical physics, radiology, medical imaging, and medical

engineering. To make a cancer patient understand the rationale of radiotherapy is essential, because through understanding how the ionizing radiation can damage the cancer cell using the most up-to-date delivery techniques, confidence in the patient can be built up helping them to fight with cancer. The chatbot can help all the radiotherapy patients by providing them not only their appointments at the hospital, but also the necessary tailor-made information related to their treatments for patient care [41]. On the other hand, for the general public, the best way to avoid cancer should be through a healthy lifestyle and cancer screening. In this event, the chatbot can be used to promote a healthy life to avoid cancer and some cancer-screening programs. The general public will be benefited by interacting with the chatbot to acquire their needed information quickly and accurately. It is because the chatbot is humanlike mimicking a cancer professional to educate the users, whether they are cancer patients or general public [42].

An AI-assisted chatbot is a two-way communication system using a real humanlike interface. Through a warm and human touch-like communication between the chatbot and user, a friendship can be built up making the user believe the chatbot and hence makes it easier to listen to it for accepting information. It should be noted that cancer often affects seniors. Senior cancer patients

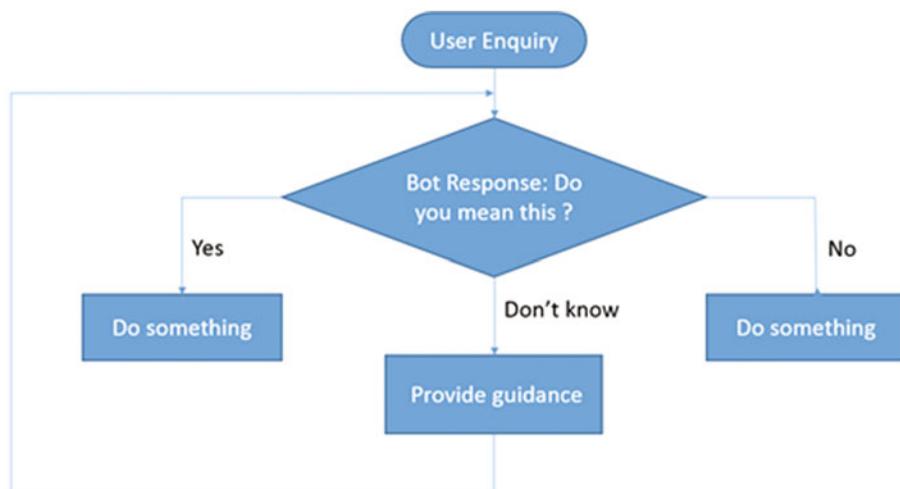


Fig. 5 The simple workflow of a chatbot

could not keep pace with the most updated computer technologies such as YouTube, Facebook, Twitter, and Instagram. They would need a relatively longer time for communication and to listen to and understand the information. The current web-based system can only provide information as per the selection made by the user on a tablet. The system cannot actually “talk” to the user like a chatbot. An AI-assisted chatbot for cancer patient is able to communicate with the user to start a conversation. Through talking with the user, the chatbot understands the background and their need. The AI system will then deliver the necessary information to the user as per their need. This dissemination of knowledge may be simple but useful, such as finding out a treatment room for appointment, the way to the washroom, patient care information, or some cancer-preventive strategies. The chatbot can solve this problem because it can predict the background and general information of the user, and find out the best way to communicate with them. The chatbot can ensure the most decent information to be acquired by the user in a warm and friendly manner.

Conclusion and Future Prospective

AI is a powerful digital tool that can benefit the advancement and automation of radiotherapy, leading to better resource allocation, more effective workflow, more accurate and faster decision-making, better patient care, enhancements of treatment plan QA, and radiation dose delivery. To date, it is seen that AI has been implemented in various components in the radiotherapy chain such as treatment planning, plan evaluation/QA, dose delivery, and patient education/care. With different AI-assisted radiotherapy processes being developed and commercialized, such as machine learning treatment planning, AI-assisted QA, AI-assisted diagnostic medical imaging, machine learning leaf sequencing for dose delivery, AI-assisted chatbot for cancer patient care, and clinical decision support and outcomes prediction, it is expected that cancer patients will be benefitted. However, there are still challenges in the application of AI in radiotherapy. These

include lack of personnel who know both AI/machine learning and radiotherapy, collaboration among multiple institutions for data privacy, security and ownership, and supports from the vendors and manufacturers for standardization of the dataset and tools and verification of results. Future directions and prospectives include focusing on the studies of data allocation and storage, because data volume will continue to grow in future with accumulation of more patient cases. Moreover, it is important to study big data management and analytics for machine learning. This involves the data-filtering technique in data-mining. Data privacy, computer security, humanlike-communication interface, and machine learning algorithm will continue to develop. Equally important is to ensure that users who do not have knowledge on the AI algorithm can still master the virtual machine easily to produce useful results.

References

1. Chow JCL. Internet-based computer technology on radiotherapy. *Rep Pract Oncol Radiother.* 2017;22: 455–62.
2. Chow JCL. Application of cloud computing in preclinical treatment planning. *Int J Comput Res.* 2015;22(3): 209–22.
3. Chow JCL. A performance evaluation on Monte Carlo simulation for radiation dosimetry using cell processor. *J Comput Methods Sci Eng.* 2011;11:1–12.
4. Siddique S, Chow JCL. Artificial intelligence in radiotherapy. *Rep Pract Oncol Radiother.* 2020;25:656–66.
5. Wang C, Zhu X, Hong JC, Zheng D. Artificial intelligence in radiotherapy treatment planning: present and future. *Technol Cancer Res Treat.* 2019;18: 1533033819873922.
6. Thompson RF, Valdes G, Fuller CD, Carpenter CM, Morin O, Aneja S, Lindsay WD, Aerts HJ, Agrimson B, Deville C Jr, Rosenthal SA. Artificial intelligence in radiation oncology: a specialty-wide disruptive transformation? *Radiother Oncol.* 2018;129(3): 421–6.
7. Alfonso JC, Herrero MA, Nunez L. A dose-volume histogram based decision-support system for dosimetric comparison of radiotherapy treatment plans. *Radiat Oncol.* 2015;10(1):263.
8. Chow JCL, Jiang R, X Lu. Evaluation of plan optimizers in prostate VMAT using the dose distribution index. *J Radiother Pract.* 2019;18(4):323–8.
9. Pearse J, Chow JCL. An Internet of Things App for monitor unit calculation in superficial and orthovoltage skin therapy. *IOP SciNotes.* 2020;1:014002.

10. Chow JCL, Grigorov GN, Yazdani N. SWIMRT: a graphical user interface using sliding window algorithm to construct fluence map machine file. *J Appl Clin Med Phys.* 2006;7:69–85.
11. Dharwadkar R, Deshpande NA. A medical ChatBot. *Int J Comput Trends Technol (IJCTT).* 2018;60(1):41–5.
12. Abdul-Kader SA, Woods JC. Survey on chatbot design techniques in speech conversation systems. *Int J Adv Comput Sci Appl.* 2015;6(7):72–80.
13. Mitchell RS, Michalski JG, Carbonell TM. An artificial intelligence approach. Berlin: Springer; 2013.
14. Jakhar D, Kaur I. Artificial intelligence, machine learning and deep learning: definitions and differences. *Clin Exp Dermatol.* 2020;45(1):131–2.
15. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science.* 2015;349 (6245):255–60.
16. Premalatha J, Ravichandran KS. Novel approaches for diagnosing melanoma skin lesions through supervised and deep learning algorithms. *J Med Syst.* 2016;40(4):96.
17. Albawi S, Mohammed TA, Al-Zawi S. Understanding of a convolutional neural network. In: 2017 international conference on engineering and technology (ICET), 2017 Aug 21. IEEE; 2017. p. 1–6.
18. Staffurth J. A review of the clinical evidence for intensity-modulated radiotherapy. *Clin Oncol.* 2010;22(8):643–57.
19. Chow JC. Recent progress in Monte Carlo simulation on gold nanoparticle radiosensitization. *AIMS Biophys.* 2018;5(4):231–44.
20. Lu L. Dose calculation algorithms in external beam photon radiation therapy. *Int J Cancer Ther Oncol.* 2014;1(2):01025.
21. McIntosh C, Purdie TG. Voxel-based dose prediction with multi-patient atlas selection for automated radiotherapy treatment planning. *Phys Med Biol.* 2016;62 (2):415.
22. Kim H, Jung J, Kim J, Cho B, Kwak J, Jang JY, Lee SW, Lee JG, Yoon SM. Abdominal multi-organ auto-segmentation using 3D-patch-based deep convolutional neural network. *Sci Rep.* 2020;10(1):1–9.
23. Wong J, Fong A, McVicar N, Smith S, Giambattista J, Wells D, Kolbeck C, Giambattista J, Gondara L, Alexander A. Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. *Radiother Oncol.* 2020;144:152–8.
24. Wang M, Zhang Q, Lam S, Cai J, Yang R. A review on application of deep learning algorithms in external beam radiotherapy automated treatment planning. *Front Oncol.* 2020;10:580919.
25. Amit G, Purdie TG, Levinstein A, Hope AJ, Lindsay P, Jaffray DA, Pekar V. Automatic learning-based selection of beam angles in radiation therapy of lung cancer. In: 2014 IEEE 11th international symposium on biomedical imaging (ISBI), 2014 Apr 29. IEEE; 2014. p. 230–3.
26. Chow JCL, Jiang R, Kiciak A. Dose-volume consistency and radiobiological characterization between prostate IMRT and VMAT plans. *Int J Cancer Ther Oncol.* 2016;4(4):447.
27. Chow JCL, Markel D, Jiang R. Dose-volume histogram analysis in radiotherapy using the Gaussian error function. *Med Phys.* 2008;35:1398–402.
28. Chan MF, Witztum A, Valdes G. Integration of AI and machine learning in radiotherapy QA. *Front Artif Intell.* 2020;3:76.
29. Markel D, Alasti H, Chow JCL. Dosimetric correction for a 4D-computed tomography dataset using the free-form deformation algorithm. *J Phys Conf Ser.* 2012;385:012001.
30. Jia X, Yan H, Gu X, Jiang B. Monte Carlo simulation for patient-specific CT/CBCT imaging dose calculation. *Phys Med Biol.* 2012;57:577–90.
31. Isa M, Jiang R, Kiciak A, Rehman J, Afzal M, Chow JCL. Dosimetric and radiobiological characterizations of prostate IMRT and VMAT: a single-institution review of 90 cases. *J Med Phys.* 2016;41:162–8.
32. Chow JCL, Jiang R. Dose-volume and radiobiological dependence on the calculation grid size in prostate VMAT planning. *Med Dosim.* 2018;43:383–9.
33. Ng F, Jiang R, Chow JCL. Predicting treatment planning evaluation parameter using artificial intelligence and machine learning. *IOP SciNotes.* 2020;1:014003.
34. Ren L, Ma Y, Shi H, Chen X. Overview of machine learning algorithms. In: Signal and information processing, networking and computers. Singapore: Springer; 2020. p. 672–8.
35. Chow JCL, Jiang R, Lu X. Dosimetric and radiobiological comparison of prostate VMAT plans optimized using the photon and progressive resolution algorithm. *Med Dosim.* 2020;45(1):14–8.
36. Chow JCL, Girgorov GN. Measurement for the MLC leaf velocity profile by considering the leaf leakage using radiographic film. *Phys Med Biol.* 2006;51: N299–306.
37. Jia J, Hui L, Chow JC. A leaf sequencing algorithm for multileaf collimator in intensity modulated radiotherapy. *Rep Radiother Oncol.* 2015;2(4):e4922.
38. Jing J, Lin H, Chow JCL. A novel computer graphical user interface for MLC leaf sequencing based on the shape optimization technique. *Med Phys.* 2020;47(6): e684.
39. Siddique S, Chow JCL. Machine learning in healthcare communication. *Encyclopaedia;* 2012;1(1):220–239.
40. Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: a mixed-methods study. *Digital Health.* 2019;5:2055207619871808.
41. Xu L, Chow JCL. Chatbot for cancer therapy and patient care using artificial intelligence and machine learning. Submitted to JMIR Bioinformatics and Biotechnology, 2021.
42. Chow JCL, Li K. A Chatbot with characterization on radiotherapy using artificial intelligence and machine learning. In: Proceedings in cancer AI and big data: success through global collaboration PMICEC; 2020. p. 32.



Deep Learning in Mammography Breast Cancer Detection

92

Richa Agarwal, Moi Hoon Yap, Md. Kamrul Hasan,
Reyer Zwiggelaar, and Robert Martí

Contents

Introduction	1288
Datasets	1288
MIAS	1289
DDSM	1289
INbreast	1289
OPTIMAM Mammography Image Database (OMI-DB)	1289
BCDR	1289
Deep Learning Methods	1291
Performance Metrics	1294
Discussion and Future Challenges	1295
Conclusion	1297
References	1297

Abstract

Breast cancer incidence has increased in the past decades. Extensive efforts are being made for early detection to reduce the mortality rate. As one of the leading field in artificial intelligence, deep learning algorithms have been widely used in breast cancer research in recent years, ranging from detection, segmentation to classification. To provide insights and development in this field, we review and summarize the popular datasets and deep learning methods used in breast cancer detection, focusing on mammography. We provide a summary of the state-of-the-art deep learning methods in mammography breast cancer detection and its performance. We discuss the challenges in

R. Agarwal · R. Martí (✉)
Computer Vision and Robotics Institute, University of
Girona, Girona, Spain
e-mail: robert.marti@udg.edu

M. H. Yap
Manchester Metropolitan University, Manchester, UK
e-mail: M.Yap@mmu.ac.uk

M. K. Hasan
Bangladesh University of Engineering and Technology,
Dhaka, Bangladesh
e-mail: khasan@eee.buet.ac.bd

R. Zwiggelaar
Aberystwyth University, Aberystwyth, UK
e-mail: rzz@aber.ac.uk

breast cancer detection and provide some new insights to advance the field.

Introduction

Breast cancer is one of the leading death causes of women. Over the past years, researchers proposed computerized techniques with the intention to aid the computer-aided diagnosis. However, many did not share their datasets and codes, which have become the main barriers in advancing the field. With the advancements in computer technology and data sciences, there has been a lot of research effort in exploring deep learning methods for various tasks. Deep learning methods based on Convolutional Neural Networks (CNNs) have also gained importance in the field of medical image analysis and the efforts are laid to develop modern computer aided detection systems.

This chapter provides an overview of publicly available mammography datasets and its most commonly used deep learning methods. For deep learning methods, we review the state-of-the-art deep learning algorithms for mammograms, compare the performance of the deep learning methods, and explore recent advancements in CNN to facilitate the development of automated deep learning method in detection of breast lesions. We discuss the challenges and future directions of deep learning in breast cancer detection.

Datasets

One of the major barriers in medical image analysis is the limited availability of datasets. Despite recent initiatives in sharing large-scale dataset in natural images [41], public datasets in the field of medical imaging and breast cancer in particular are still scarce or nonexistent. This can be explained by the difficulties to obtain ethical approvals to share the patients' data but also to the need of obtaining accurate data annotations and ground truth, which often requires gathering complex sources of information such as clinical data (i.e., electronic patient records), tests (i.e.,

biopsy results), and specialists (i.e., radiologists), therefore being both complex technically and time-consuming.

Nevertheless, public datasets need to be encouraged and developed as they allow the advance on the state-of-the-art methods on a particular field, providing a common platform to train machine learning methods but also, and more importantly, to compare and evaluate different methods. However, in order to be useful, these public datasets need to be of clinical relevance for the particular medical field or topic and need to include ground truth as accurate as possible, minimizing the effects of intra and inter reader variability.

There is a number of publicly available datasets for mammographic images: DDSM [33], CBIS-DDSM [44], INbreast [52], MIAS [63], and BCDR [29]. In addition, there are other initiatives that, although not publicly available, provide access to large datasets of mammographic datasets including clinical and patient data, subject to application. Those projects include the OMI-DB project in the UK [19] or the recent Cohort of Screen-Aged Women (CSAW) dataset from the Stockholm region in Sweden [20]. In addition to obvious aspects such as the number of patients, imaging views (CC, MLO), variability of lesions (benign and malign), and clinical information, an important aspect about a mammographic dataset is the nature of the images themselves, as older datasets [33, 44, 63] contain digitized screen film mammograms (SFM), where more recent databases include full field digital mammograms (FFDM) and, in some cases, even mammograms in raw format (also known as for processing) in addition to the mammograms post-processed for visualization (also known as for presentation). This chapter does not cover other modalities closely related to mammography such as Digital Breast Tomosynthesis (DBT), magnetic resonance imaging, or ultrasound. In those modalities, the number of publicly available datasets is even less common compared to mammography, although some initiatives can be found, such as the MRI dataset used in [30], the breast ultrasound in [73], or the recent DBT lesion detection challenge in [6].

MIAS

The Mammographic Image Analysis Society (MIAS) dataset [63] was one of the first public datasets. It was published in 1994 and contains 322 SFM, initially digitized at 50 microns but resized to 200 microns with a common resolution of 1024×1024 pixels. The database includes 209 normal mammograms and 113 abnormalities with radiologist annotations, indicating the type of abnormality (calcifications, well-defined/circumscribed, spiculated, ill-defined masses, and architectural distortions and asymmetries) and the position and extent of the lesion. The database also includes a classification of each image according to their density pattern using three classes: fatty, fatty-glandular, and dense-glandular.

DDSM

The Digital Database for Screening Mammography (DDSM) dataset contains approximately 2,500 studies. It contains SFM compressed with lossless JPEG encoding. The Curated Breast Imaging Subset of DDSM (CBIS-DDSM) [43] is a newer version of the mammographic dataset, containing a subset of the original DDSM images in the standard DICOM format. It was curated by a trained mammographer. The dataset can be downloaded from CBIS-DDSM website [17]. It contains 3,061 mammograms of 1,597 cases. In total there are 1,698 masses in 1,592 images from 891 cases which includes both CC and MLO views for most of the screened breasts. Figure 1 shows two mammographic views of the same case (CC and MLO) from the dataset.

INbreast

The INbreast [52] dataset is composed of FFDM acquired using a Siemens MammoNovation mammography system (Siemens Healthineers, Erlangen, Germany). The images were acquired from 115 cases with CC and MLO breast views, leading to a total of 410 images available in

DICOM format. From these, a total of 116 masses can be found in 107 mammograms from 50 cases. Several types of lesions (masses, calcifications, asymmetries, and distortions) are included. Accurate contours made by specialists are also provided in XML format. Figure 2 shows two views of an FFDM (CC and MLO) of the same case in the dataset.

OPTIMAM Mammography Image Database (OMI-DB)

The OMI-DB [32] is an extensive mammography image database of over 145,000 cases (over 2.4 million images) comprised of unprocessed and processed FFDM from the NHS Breast Screening Programme of the United Kingdom, which also contains expert's determined GT and associated clinical data linked to the images. As this is a relatively new dataset, Agarwal et al. [2] were the first to implement a deep learning algorithm on a subset of this dataset, comprising of 4,750 cases. These cases are images captured from four different manufacturers including Hologic, Philips, General Electric (GE), and Siemens, containing a total of 2,419 cases with 4,217 masses and 946 cases without any mass. Figure 3 shows sample FFDMs from the scanners of different manufacturers in the OMI-DB dataset. This is a much richer dataset as it also contains clinical, surgical (i.e., biopsy related) and pathological information about each patient, in addition to the mammographic images.

BCDR

The Breast Cancer Digital Repository (BCDR) [29] dataset contains 1,734 patients with mammography, clinical history, lesion segmentation, and various precomputed image-based features. The dataset is divided into two different repositories: a digitized SFM Repository and a FFDM repository both obtained from the Centro Hospitalar São João, at University of Porto, Portugal. BCDR provides normal and annotated patients cases of breast cancer including

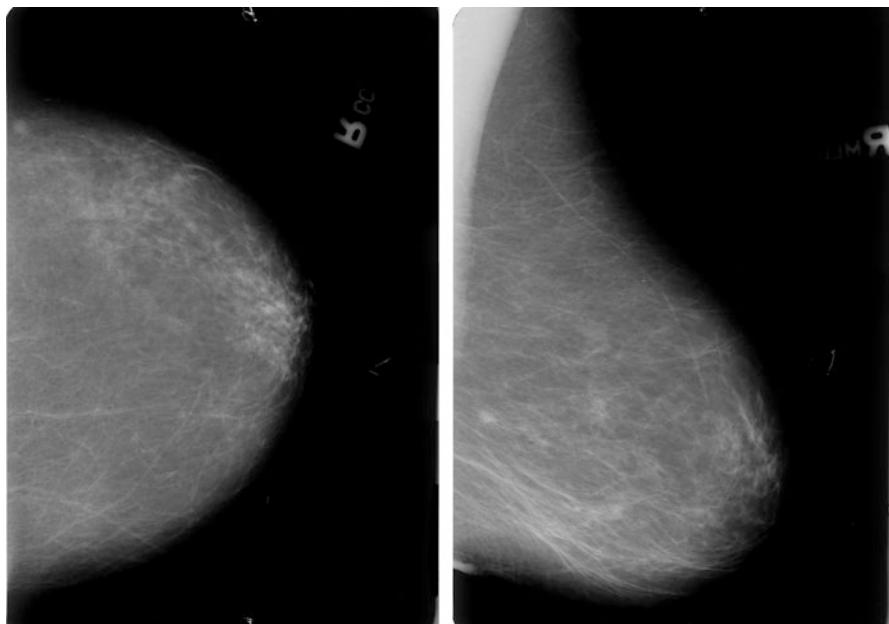


Fig. 1 Sample SFM from CBIS-DDSM [43] showing two different views of the same case, left: CC and right: MLO

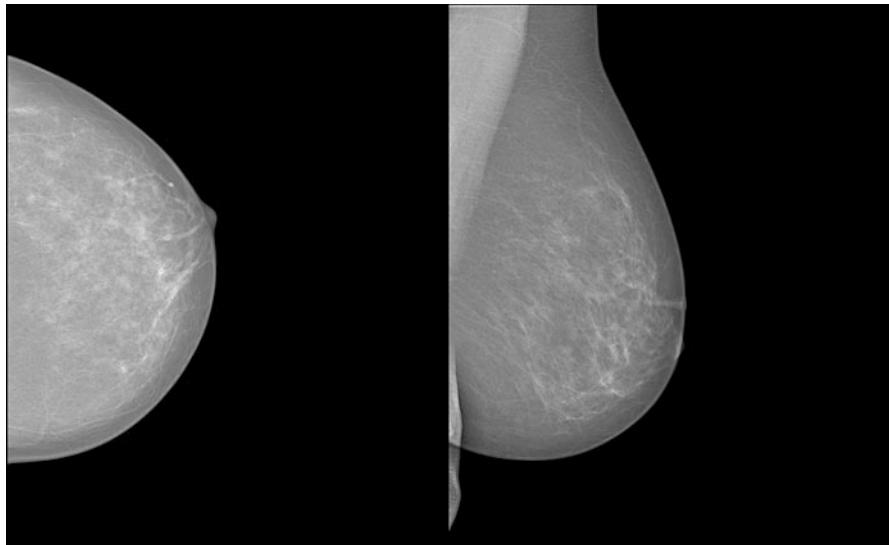


Fig. 2 Sample FFDM from INbreast [52] showing two different views of the same case, left: CC and right: MLO

mammography lesions outlines, anomalies observed by radiologists, precomputed image-based descriptors, as well as related clinical data. The digitized database is composed of 1010 patients (998 female and 12 male) with ages between 20 and 90 years old. From those patients,

there are 1,125 studies with 3,703 mediolateral oblique (MLO) and craniocaudal (CC) images and 1,044 lesions manually segmented and classified using the BIRADS standard for malignancy assessment. Images were digitized and converted to TIFF format with a resolution of 720×1168

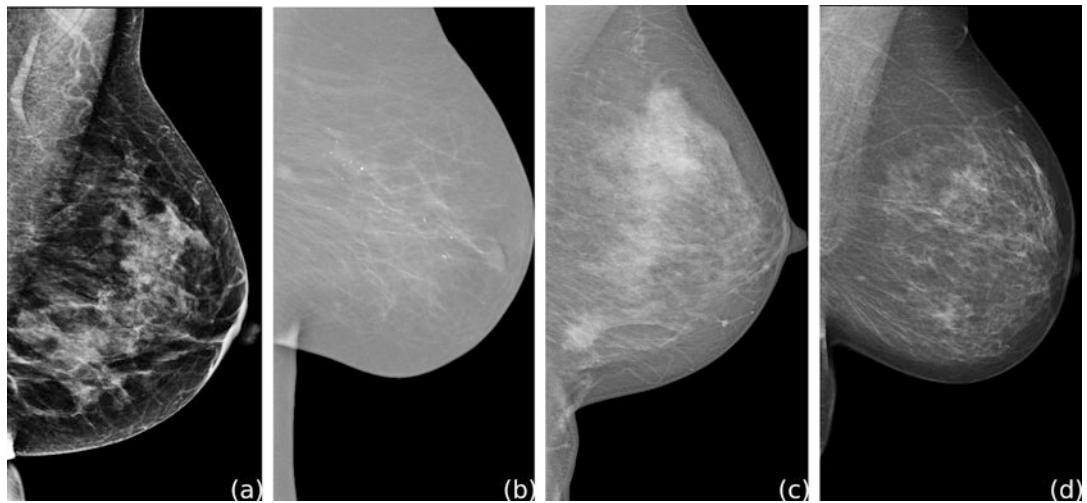


Fig. 3 Sample FFDM from OMI-DB dataset from different manufacturers. Here mammograms are from (a) Hologic, (b) GE, (c) Siemens, and (d) Philips

pixels and a bit depth of 8 bits per pixel. The FFDM dataset is composed of 724 patients (ages between 27 and 90 years old) with 1042 studies, 3612 MLO/CC images and 452 lesions with manual segmentation. The resolution of the images is 3328×4084 or 2560×3328 pixels (depending on the acquisition parameters) and saved as TIFF format using 14 bits per pixel.

Deep Learning Methods

Following the recent developments in computer technology and data science, there has been a lot of interest in exploring deep learning methods for breast cancer [1, 37, 70, 73]. The term “Deep Learning” can be defined as a subtopic of machine learning methods that learn data representations using multiple levels of representation [42]. They are obtained by composing simple but nonlinear models that transform the representation from one level (starting with the raw input) into increasing levels of representation. Deep learning strategies have recently gained a lot of interest in various fields including object detection [12, 27, 42, 57, 65], image recognition [24, 27, 41, 57, 64–66], natural language processing [13, 18], and speech recognition [34, 50]. In Deep Learning, Convolutional

Neural Network (CNN) is most commonly used to analyze images.

This section outlines the recent deep learning methods for breast cancer in mammography. Table 1 presents a summary of the state-of-the-art deep learning methods on mammography mass detection by year, dataset used, image resolution, and performance measures.

Several authors have proposed the use of traditional machine learning and content-based image retrieval techniques to classify masses and microcalcifications [25, 54]. The feature exploitation of deep learning frameworks in the field of breast imaging has been relatively limited, as only a small number of public datasets are available (as seen in the previous section).

In this regard, one of the first works was the work of Dhungel et al. [21] in 2015, who proposed a multiscale deep belief network (m-DBN) classifier, followed by a cascade of Region-Based CNN (R-CNN) and cascades of random forest classifiers for mass detection, obtaining True Positive Rate (TPR) of 0.96 ± 0.03 at 1.2 False Positives per Image (FPI) on INbreast and 0.75 at 4.8 FPI on DDSM-BCRP dataset. Later, Carneiro et al. [15] proposed the use of CNN models pretrained using a computer vision database (ImageNet) for classifying benign and

Table 1 Summary of the state-of-the-art deep learning methods on mass detection in mammography

Author (Year)	Dataset (Images)	Method	Mammogram size	AUROC	TPR at FPI
Kozegar et al. [40] (2013)	mini-MIAS (330)	Iterated segmentation	512 × 512	n/a	0.91 at 4.8
	INbreast (116)				0.87 at 3.67
Carneiro et al. [15] (2015)	INbreast (410)	Fast CNN (CNN-F) [16]	264 × 264	0.91 ± 0.05	n/a
	DDSM (680)			0.97 ± 0.03	
Dhungel et al. [21] (2015)	INbreast (410)	m-DBN + R-CNN + RF	264 × 264	n/a	0.96 ± 0.03 at 1.2
	DDSM-BCRP (316)				0.75 at 4.8
Lotter et al. [48] (2017)	DDSM (10480)	InceptionV3	patch (256 × 256)	0.77 ± 0.03	n/a
		wide ResNet [74]		0.92 ± 0.02	
Dhungel et al. [22] (2017)	INbreast (410)	Cascade R-CNN + RF with hypothesis refinement	264 × 264	n/a	0.90 ± 0.02 at 1.3
Becker et al. [11] (2017)	BCDR (70)	ANN using ViDi software	high resolution	0.79	n/a
Kooi et al. [39] (2017)	internal (45000)	Downscaled VGG	patch (250 × 250)	0.929	n/a
Akselrod-Ballin et al. [4] (2017)	internal (850)	Modified Faster R-CNN	grid (800 × 800)	0.78	n/a
Akselrod-Ballin et al. [5] (2017)	internal (750)	Modified Faster R-CNN	grid (800 × 800)	n/a	0.9 at 1.0
	INbreast (100)				0.93 at 0.56
Ribli et al. [58] (2018)	INbreast (n/a)	Faster R-CNN	2100 × 1700	0.95	0.90 at 0.3
Jung et al. [38] (2018)	GURO (222)	RetinaNet	n/a	n/a	0.98 ± 0.02 at 1.3
	INbreast (410)				0.94 ± 0.05 at 1.3
Morrel et al. [53] (2018)	Dream challenge (13,000)	R-FCN/DCN	2,545 × 2,545	0.8667	n/a
Al-masni et al. [8] (2018)	DDSM (600)	YOLO	448 × 448	0.877	n/a
Al-antari et al. [7] (2018)	INbreast (410)	YOLO	448 × 448	0.948	n/a
Agarwal et al. [1] (2019)	INbreast (410)	Pat CNN	224 × 224	n/a	0.98 ± 0.02 at 1.67
Agarwal et al. [2] (2020)	INbreast (410)	Faster-RCNN	256 × 256	0.90	0.95 ± 0.03 at 1.14
Shayma et al. [61] (2020)	MIAS (70)	AlexNet	227 × 227	0.989	n/a
Aly et al. [9] (2020)	INbreast (410)	YOLOV3	832 × 832	0.98	0.92 at 0.086

malignant lesions obtaining an Area Under Receiver Operating Curve (AUROC) of 0.91 ± 0.05 on INbreast and 0.97 ± 0.03 on DDSM dataset.

In 2017, Lotter et al. [48] trained a patch-based CNN to classify lesions in the DDSM dataset and subsequently used a scanning-window approach to provide full mammogram classification

achieving an AUROC of 0.92 on the DDSM dataset. In the scanning-window approach, the image is partitioned into set of patches (with minimal overlap) and each patch is the input of the CNN and classified as being normal or lesion. Posteriorly, Dhungel et al. [22] used a deep learning approach for mass detection, segmentation, and classification in mammograms and evaluated on the INbreast dataset. Detection results had a TPR of 0.90 ± 0.02 at 1.3 FPI on testing set.

Regarding the use of private mammography datasets, Becker et al. [11] used a multipurpose image analysis software and an internal database from 3,228 patients to train the Artificial Neural Network (ANN). The model was then tested on the BCDR dataset [47] of 35 patients to obtain AUROC of 0.79. In other works, Kooi et al. [39] used a larger private database of 45,000 FFDM to provide a comparison between traditional mammography CAD systems relying on hand-crafted features and the CNN methods. It was shown that the CNN model trained on a patch level with a large database outperformed state-of-the-art CAD systems, obtaining an AUROC of 0.929 compared to 0.906 obtained for the CAD system.

Researchers have used Faster R-CNN in medical imaging [28, 67], but a very small amount of literature is available in the field of breast imaging. For instance, Akselrod-Ballin et al. [4] used a modified version of Faster R-CNN model to include information from the finer bottom levels during classification stage. Results were then evaluated on an internal database of 850 images to obtain AUROC of 0.78. Later, in [5], the results were evaluated on a subset of the INbreast dataset (with masses) obtaining a TPR of 0.93 at 0.56 FPI. Similarly, Ribli et al. [58] trained a Faster R-CNN on the DDSM database composed of 2,620 SFM and then evaluated the performance of the network on the INbreast database of malignant lesions, obtaining a TPR of 0.90 at 0.3 FPI and AUROC of 0.95.

Jung et al. [38] proposed a mass detection model based on RetinaNet [46] and used a new loss function called focal loss to address the problem of extreme class imbalance between foreground and background. The performance of the network was evaluated on the INbreast to obtain

the best TPR of 0.97 ± 0.05 at 3.0 FPI and 0.94 ± 0.02 at 1.3 FPI. Morrel et al. [53] presented a neural network based on region-based fully convolutional network (R-FCN) and deformable convolutional nets. The network was trained using the OMI-DB [32] dataset, achieving AUROC of 0.867 for breast-wise detection in the DREAM challenge on 13,000 images from Group Health.

Al-masni et al. [8] adopted the You Only Look Once (YOLO) deep learning method [56] for the detection and classification of masses in mammograms. The results showed an overall mass detection accuracy of 96.33% and a classification accuracy of 85.52% on the DDSM dataset. Authors showed that enhanced accuracies were obtained when using an augmented DDSM database created by rotating each mammogram multiple times. However, they used images of the same patient for both training and testing instead of splitting training and testing datasets at a patient level, thus giving potentially biased results. In other works, Al-antari et al. [7] presented a fully integrated CAD system adding lesion segmentation to the framework proposed in [8]. The results were presented on the INbreast dataset, but again the augmented dataset is used and the distribution is made based only on the images raising issue on overlap between training and testing patients.

Agarwal et al. [1] presented an automated mass detection framework using CNNs, where small regions of the mammograms (patches) were extracted using a sliding window approach and used for training different CNNs. The framework obtained results comparable to the state-of-the-art on the INbreast dataset. However, the high computational cost was a limiting factor for its clinical use. In 2020 [2], a Faster-RCNN-based framework was proposed to detect masses in the large-scale OMI-DB dataset consisting of approximately 80,000 FFDMs, obtaining a TPR of 0.93 at 0.78 FPI on FFDMs from the Hologic scanner. The framework was also used to analyze the state-of-the-art INbreast dataset, obtaining a TPR of 0.99 ± 0.03 at 1.17 FPI for malignant and 0.85 ± 0.08 at 1.0 FPI for benign masses. Figure 4 depicts the overall framework for the detection of lesions using the Faster-RCNN algorithm.

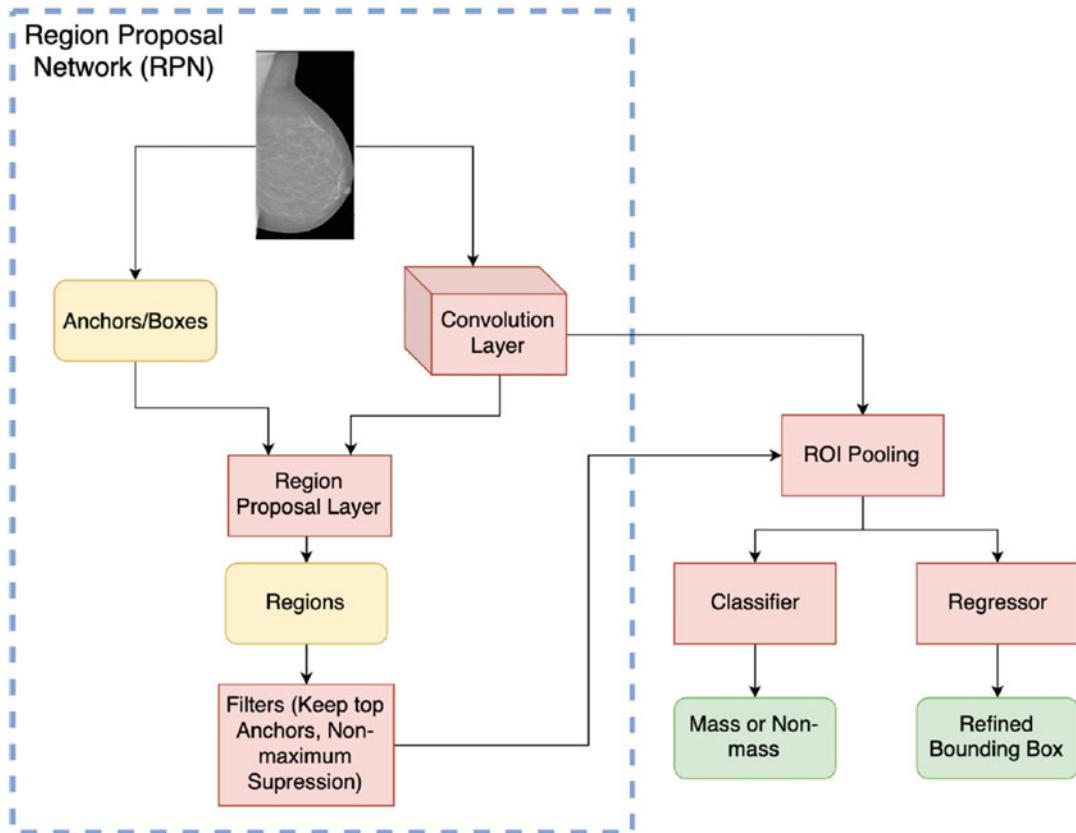


Fig. 4 Diagram of Faster-RCNN for lesion detection in mammography [2]

The latest developments in this field include the work of Shayma et al. [61] and Aly et al. [9], where a framework based on AlexNet and YOLO has been used for the classification and detection of breast masses, respectively.

Despite the limited availability of datasets, researchers have successfully implemented deep learning algorithms with data augmentation and two strategies: (i) by using image patches as inputs rather than full sized image and (ii) use the concept of transfer learning, in which the knowledge obtained by training for one task is used for another related task (also known as domain adaptation). In this regard, firstly the CNN trained on a very large dataset of natural images is adapted to classify between mass and nonmass image patches in the Screen-Film Mammogram (SFM), and secondly the newly trained CNN model is adapted to detect masses in FFDM. The following section describes some popular

performance metrics, which are commonly used to evaluate the performance of breast screening.

Performance Metrics

Different measures are commonly used for the evaluation of lesion segmentation methods. Usually the terms segmentation and detection need to be differentiated in medical image analysis. A detection method aims at obtaining the position and extent (i.e., with a bounding box if possible) of a lesion but without an accurate boundary delineation. On the other hand, a lesion segmentation method is designed to obtain an accurate outline, region, or contour of the lesion for subsequent analysis (i.e., diagnosis). Segmentation algorithms are often applied after the detection of the lesion is performed, although some methods (i.e., U-Net) provide both segmentation

and detection simultaneously. One should note that this chapter focuses on lesion detection; hence, evaluation measures for this task are discussed below.

Most measures take into account the elements of a confusion matrix: true positives and negatives (TP, TN) and false positives and negatives (FP, FN). In that sense, the definition of what is considered a TP needs to be clearly stated for the methods to be comparable. Usually the criteria of overlapping with the ground truth bounding box region is used at a given threshold criteria (i.e., > 50% of overlap is considered a TP). However, in the literature, one can find more loose definitions, such as defining a detection as a TP if the algorithm finds a lesion in an image which has been labeled as containing a lesion, regardless of its correct localization. In any case, from a given definition of TP, several evaluation measures can be derived. True Positive Fraction (TPF) and False Positives per image (FPs/image) are the most commonly used measures [23, 60, 69]:

$$TPF = \frac{\text{number of TPs}}{\text{number of actual lesions}} \quad (1)$$

$$FPs/\text{image} = \frac{\text{number of FPs}}{\text{number of images}} \quad (2)$$

TPF measures the sensitivity of the method. Some of the algorithms are capable of detecting multiple lesions while some are only capable of detecting a single lesion. The TPF allows a fair measurement as it is measuring the total detected lesions to the total number of actual lesions. Thus, if a method can detect only one lesion in an image with multiple lesions, the TPF of this methodology will be lower than the method that is capable of detecting multiple lesions.

In addition, Receiver Operating Characteristic (ROC) analysis can be used, as a graphic plot that illustrates the performance of a classifier system as a function of its discrimination threshold (i.e., probability of being a lesion). It is created by plotting the fraction of true positives out of the total actual negatives (FPR) for various thresholds. The Area Under the Curve (referred to as *AUC* or A_z) can be computed from the ROC curve

as an quantitative indicator of overall performance of a classifier. Often detection algorithms are also evaluated using free-response ROC analysis (FROC), a variation of ROC analysis, where the X-axis is replaced by the number of false positives per image. FROC gives more importance to the correct localization of the lesion rather than just analyzing the false positive rate.

In addition to TPF and FPs/image, the F-measure (the weighted harmonic mean of recall and precision) [55] is computed as:

$$F\text{-measure} = \frac{2 \times \text{TP}}{(2 \times \text{TP}) + \text{FP} + \text{FN}} \quad (3)$$

Discussion and Future Challenges

As mammography is the gold standard for breast cancer screening, there is a growing trend adopting deep learning for real-world applications. A recent study in nature [49] showed an artificial intelligence system is capable in breast cancer prediction, surpassing human performance. The work is based on a combination of OPTIMAM database and a private dataset – which makes it difficult to replicate and compare against.

In specific patient populations (i.e., young women or women with dense breasts), mammography is often nonconclusive and other modalities are needed. For instance, ultrasound is used as a complementary tool for breast screening. Breast ultrasound examination is a safe procedure as it is noninvasive and does not use ionizing radiation. Researchers in breast ultrasound are facing the similar challenges of small datasets available for research. The earlier works in using deep learning for breast ultrasound were in 2016 by [37], but particularly motivated by Yap et al. [73] in 2017 when they shared a dataset. Although this is a small dataset, it has provided the researchers a common ground to compare the performance of their algorithms. With transfer learning and inspired by the success of fully convolutional network and U-Net, researchers have experimented deep learning methods in segmentation [14, 70, 72] and region

of interests detection [71]. As breast ultrasound is widely used alongside with mammography in clinical practices, future challenge is to develop deep learning methods to work on several modalities.

In recent times, tissue stiffness-based biomarker is emerging with enormous potential for differentiating between benign and malignant breast lesions because of the fact that malignant tumor tissues are harder than normal and benign ones [31, 36]. Ultrasound elastography is the imaging modality being currently used for measuring the stiffness of the tissue along with morphology. Among the variants of elastography, strain elastography and shear-wave elastography are the two well-studied techniques for breast imaging. The diagnostic performance of deep learning-based shear-wave elastography has been demonstrated to be very promising compared to conventional ultrasound [75]. However, combining complementary information from ultrasound B-mode and elastography can provide even better diagnostic decision in computer-aided diagnosis of breast lesions [35]. In particular, elasticity imaging may help early detection of breast cancer as the changes in breast tissue elasticity may result earlier than its morphological changes [10]. Though substantial advances have been made, the elastography imaging still requires further attention for improving the image quality [3]. Fusion of mammography, ultrasound, elastography, and photo acoustic imaging may be the next generation technology for computer-aided diagnosis of breast cancer.

With the advancement of 3D technology and hardware capabilities, magnetic resonance imaging (MRI) is used as a supplemental tool for mammography and/or ultrasound. It may be used to screen women at high risk for breast cancer, evaluate the extent of cancer following diagnosis, or further evaluate abnormalities seen on mammography. Typical breast MRI is performed on a 1.5 Tesla magnet with a dedicated multichannel breast coil. In breast imaging, dynamic contrast-enhanced MRI (DCE-MRI) is used and is a noninvasive process. Contrast agents play a crucial role in DCE-MRI and should be carefully selected in order to improve accuracy in DCE-MRI examination [68]. Breast MRI is

expensive and produces false positives; therefore, it is not recommended for screening general population.

In mammography, there are some efforts over the past few years in expanding the capacity of mammogram to improve breast cancer detection. A new imaging tool, pseudo-3D digital breast tomosynthesis (DBT), was developed to reduce the masking effect of overlapping fibro-glandular tissue [26]. Although DBT is gradually being adopted, x-ray mammography is still the gold standard imaging modality used for breast cancer screening due to its fast acquisition and cost-effectiveness. Moreover, DBT is relatively new and its availability at the hospitals is limited. So, it is not yet considered to be a standard method for breast cancer screening.

The latest 3D technology proposed for breast screening is Automated Breast Ultrasound (ABUS), which produces a 3D volume of a full scan of breast tissues [62]. ABUS can overcome the issues of hand-held ultrasound screening in two aspects: operator dependent and lesion localisation. The challenge of ABUS is that the radiologists have to analyze each slice of the 3D volume to identify suspicious lesions, which is very time-consuming and laborious. Currently, researchers are working on developing deep learning methods for computer-aided detection [45, 51] to overcome this issue. However, there are no publicly available dataset of ABUS and lack of collaborative works to overcome this challenge. A larger publicly available ABUS dataset is required to improve the accuracy and the repeatability of the developed deep learning framework.

The deep learning methods have the potential to be adapted to detect lesions in 3D volumes such as ABUS and DBT, which are currently being adopted in the clinical practice. In this regard, an extensively large database composed of slices of 3D volumes would be required for training and testing of the CNN. This would require higher computing resources and may require using multiple GPUs for the purpose of training the CNN.

Another barrier for the adoption of deep learning methods in CAD is human acceptance. Developing a fully automated breast cancer detection algorithm is not meant to fully replace the

radiologists. As mammography requires a second reading or consensus opinion often in complex cases, the computer algorithm can step in to help in this process and reduce the workload and improve efficiency. For instance, in [59], authors published a study on the use of a CAD system to independently read around 50% of screening cases (2,600 cases, 650 of which contained cancer) and make radiologists focus only on the most complex and malign cases. The study showed no reduction on the overall detection performance compared to reading all cases by radiologists, but reducing their workload by almost half. Additionally, if CAD systems are able to incorporate all available modalities in the diagnosis process this will surely help to improve the accuracy. Fusion of modalities is a complex task due to the different nature of the different images and requires new methods to integrate this information. Other issues include trust in AI, data protection, accountability, which have open the door for researchers and policy makers to work together on the explainability of AI and establish new law/policy for using AI effectively and safely.

Conclusion

This chapter has presented an overview of the development of deep learning in mammography breast cancer detection. First we review the datasets, and then we summarize the performance of deep learning methods in mammography breast cancer detection. The emergence of data-driven deep learning has revealed the important of data sharing in medical imaging, more than ever, which has been a known challenge for many over the past decade. Although recent research shows promising results in using deep learning on mammography, we recommend to explore the potential fusion of multimodalities to improve the performance of breast screening. Finally, we discuss the challenges and provide some new insights for future work to advance the field.

Acknowledgments This work was partly funded by the research project ICEBERG: Image Computing for Enhancing Breast Cancer Radiomics (RTI2018-096333-B-I00)

from the Spanish Ministry of Science, Innovation and Universities and the GCRF QR 2019/20 (316258) funded by Research England Development Fund.

References

1. Agarwal R, Diaz O, Lladó X, Yap MH, Martí R. Automatic mass detection in mammograms using deep convolutional neural networks. *J Med Imaging*. 2019;6(3):031409.
2. Agarwal R, Diaz O, Yap MH, Lladó X, Martí R. Deep learning for mass detection in full field digital mammograms. *Comput Biol Med*. 2020;121:103774.
3. Ahmed S, Kamal U, Hasan MK. Dswe-net: a deep learning approach for shear wave elastography and lesion segmentation using single push acoustic radiation force. *Ultrasonics*. 2020;110:106283.
4. Akselrod-Ballin A, Karlinsky L, Alpert S, Hashoul S, Ben-Ari R, Barkan E. A CNN based method for automatic mass detection and classification in mammograms. *Comp Methods Biomech Biomed Eng Imag Visualiz*. 2019;7(3):242–9.
5. Akselrod-Ballin A, Karlinsky L, Hazan A, Bakalo R, Horesh AB, Shoshan Y, Barkan E. Deep learning for automatic detection of abnormal findings in breast mammography. In: Deep learning in medical image analysis and multimodal learning for clinical decision support. Québec City, QC, Canada: Springer; 2017. p. 321–9.
6. Mazurowski M, et al. MM. SPIE-AAPM-NCI DAIR digital breast tomosynthesis lesion detection challenge (dbtex). <http://spie-aapm-nci-dair.westus2.cloudapp.azure.com/competitions/>
7. Al-antari MA, Al-masni MA, Choi MT, Han SM, Kim TS. A fully integrated computer-aided diagnosis system for digital x-ray mammograms via deep learning detection, segmentation, and classification. *Int J Med Inform*. 2018;117:44–54.
8. Al-masni MA, Al-antari MA, Park JM, Gi G, Kim TY, Rivera P, Valarezo E, Choi MT, Han SM, Kim TS. Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLOs-based CAD system. *Comput Methods Prog Biomed*. 2018;157:85–94.
9. Aly GH, Marey M, El-Sayed SA, Tolba MF. Yolo based breast masses detection and classification in full-field digital mammograms. *Comput Methods Prog Biomed*. 2020;200:105823.
10. Barr RG, Nakashima K, Amy D, Cosgrove D, Farrokh A, Schafer F, Bamber JC, Castera L, Choi BI, Chou YH, et al. Wfumb guidelines and recommendations for clinical use of ultrasound elastography: part 2: breast. *Ultrasound Med Biol*. 2015;41(5):1148–60.
11. Becker A, Marcon M, Ghafoor S, Wurnig MC, Frauenfelder T, Boss A. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. *Investig Radiol*. 2017;52(7):434–40.

12. Birdwell R, Ikeda D, O'Shaughnessy K, Sickles E. Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection. *Radiology*. 2001;219:192–202.
13. Bordes A, Chopra S, Weston J. Question answering with subgraph embeddings. In: Proceedings of the 2014 conference on empirical methods in natural language processing. Doha, Qatar: EMNLP; 2014. p. 615–20.
14. Byra M, Jarosik P, Szubert A, Galperin M, Ojeda-Fournier H, Olson L, O'Boyle M, Comstock C, Andre M. Breast mass segmentation in ultrasound with selective kernel u-net convolutional neural network. *Biomed Signal Process Control*. 2020;61: 102027.
15. Carneiro G, Nascimento J, Bradley AP. Unregistered multiview mammogram analysis with pre-trained deep learning models. In: International conference on medical image computing and computer-assisted intervention. Cham: Springer; 2015. p. 652–60.
16. Chatfield K, Simonyan K, Vedaldi A, Zisserman A. Return of the devil in the details: delving deep into convolutional nets. CoRR [abs/1405.3531](https://arxiv.org/abs/1405.3531). (2014). <http://arxiv.org/abs/1405.3531>
17. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, et al. The cancer imaging archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging*. 2013;26(6):1045–57.
18. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. *J Mach Learn Res*. 2011;12: 2493–537.
19. Halling-Brown MD, Looney PT, Patel MN, Warren LM, Mackenzie A, Young KC. The oncology medical image database (OMI-DB). In: Law MY, Cook TS, editors. Medical imaging 2014: PACS and imaging informatics: next generation and innovations, vol. 9039. San Diego, California, United States: SPIE; 2014. p. 903906.
20. Dembrower K, Lindholm P, Strand F. A multi-million mammography image dataset and population-based screening cohort for the training and evaluation of deep neural networks – the cohort of screen-aged women (CSAW). *J Digit Imaging*. 2020;33(2):408–13.
21. Dhungel N, Carneiro G, Bradley AP. Automated mass detection in mammograms using cascaded deep learning and random forests. In: International conference on digital image computing: techniques and applications (DICTA). Adelaide, SA, Australia: IEEE; 2015. p. 1–8.
22. Dhungel N, Carneiro G, Bradley AP. A deep learning approach for the analysis of masses in mammograms with minimal user intervention. *Med Image Anal*. 2017;37:114–28.
23. Drukker K, Giger ML, Horsch K, Kupinski MA, Vyborny CJ, Mendelson EB. Computerized lesion detection on breast ultrasound. *Med Phys*. 2002;29(7):1438–46.
24. Farabet C, Couprie C, Najman L, LeCun Y. Learning hierarchical features for scene labeling. *IEEE Trans Pattern Anal Mach Intell*. 2013;35(8):1915–29. <https://doi.org/10.1109/TPAMI.2012.231>.
25. Giger ML, Karssemeijer N, Schnabel JA. Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer. *Annu Rev Biomed Eng*. 2013;15: 327–57.
26. Gilbert FJ, Tucker L, Young KC. Digital breast tomosynthesis (dbt): a review of the evidence for use as a screening tool. *Clin Radiol*. 2016;71(2):141–50.
27. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the conference on computer vision and pattern recognition. Columbus, OH, USA: IEEE; 2014. p. 580–7.
28. Goyal M, Reeves ND, Rajbhandari S, Yap MH. Robust methods for real-time diabetic foot ulcer detection and localization on mobile devices. *IEEE J Biomed Health Inform*. 2018;23(4):1730–41.
29. Guevara Lopez MA, Posada N, Moura D, Pollán R, Franco-Valiente J, Ortega C, Del Solar M, Díaz-Herrero G, Ramos I, Loureiro J, Fernandes T, Araújo B. Bcdr: A breast cancer digital repository. In: 15th international conference on experimental mechanics. FEUP-EURASEM-APAET, Porto/Portugal. 2012. p. 1065–6.
30. Guo W, Li H, Zhu Y, Lan L, Yang S, Drukker K, Morris E, Burnside E, Whitman G, Giger ML, Ji Y, TCGA Breast Phenotype Research Group. Prediction of clinical phenotypes in invasive breast carcinomas from the integration of radiomics and genomics data. *J Med Imaging*. 2015;2(4):1–12. <https://doi.org/10.1117/1.JMI.2.4.041007>.
31. Hall TJ, Zhu Y, Spalding CS. In vivo real-time free-hand palpation imaging. *Ultrasound Med Biol*. 2003;29(3):427–35.
32. Halling-Brown MD, Looney PT, Patel MN, Warren LM, Mackenzie A, Young KC. The oncology medical image database (omi-db). 2014. <https://doi.org/10.1117/12.2041674>
33. Heath M, Bowyer K, Kopans D, Moore R, Kegelmeyer P. The digital database for screening mammography. Proceedings of the fourth international workshop on digital mammography. Medical Physics Publishing, Madison, WI, USA. 2000.
34. Hinton G, Deng L, Yu D, Dahl GE, Mohamed A, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath TN. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process Mag*. 2012;29(6):82–97. <https://doi.org/10.1109/MSP.2012.2205597>.
35. Hossen Z, Abrar MA, Ara SR, Hasan MK. Rate-ipath: on the design of integrated ultrasonic biomarkers for breast cancer detection. *Biomed Signal Process Control*. 2020;62:102053.
36. Hoyt K, Castaneda B, Zhang M, Nigwekar P, di Sant'Agnese PA, Joseph JV, Strang J, Rubens DJ, Parker KJ. Tissue elasticity properties as biomarkers

- for prostate cancer. *Cancer Biomark.* 2008;4(4–5):213–25.
37. Huynh B, Drukier K, Giger M. Mo-de-207b-06: computer-aided diagnosis of breast ultrasound images using transfer learning from deep convolutional neural networks. *Med Phys.* 2016;43(6):3705.
38. Jung H, Kim B, Lee I, Yoo M, Lee J, Ham S, Woo O, Kang J. Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network. *PLoS One.* 2018;13(9):e0203355.
39. Kooi T, Litjens G, van Ginneken B, Gubern-Mérida A, Sánchez CI, Mann R, den Heeten A, Karssemeijer N. Large scale deep learning for computer aided detection of mammographic lesions. *Med Image Anal.* 2017;35:303–12.
40. Kozegar E, Soryani M, Minaei B, Domingues I. Assessment of a novel mass detection algorithm in mammograms. *J Cancer Res Ther.* 2013;9(4):592.
41. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. Cambridge, MA: MIT Press; 2012. p. 1097–105.
42. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521(7553):436–44. <https://doi.org/10.1038/nature14539>.
43. Lee RS, Gimenez F, Hoogi A, Miyake KK, Gorovoy M, Rubin DL. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data.* 2017;4:170177.
44. Lee RS, Gimenez F, Hoogi A, Miyake KK, Gorovoy M, Rubin DL. Data descriptor: a curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data.* 2017;4. <http://www.nature.com/scientificdata>
45. Li Y, Wu W, Chen H, Cheng L, Wang S. 3d tumor detection in automated breast ultrasound using deep convolutional neural network. *Med Phys.* 2020;47:5669–5680.
46. Lin T, Goyal P, Girshick R, He K, Dollár P. Focal Loss for Dense Object Detection. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 42(2):318–327. <https://doi.org/10.1109/TPAMI.2018.2858826>.
47. Guevara Lopez MA, Posada N, Moura D, Pollán R, Franco-Valiente J, Ortega C, Del Solar M, Díaz-Herrero G, Ramos I, Loureiro J, Fernandes T, Araújo B. BCDR: A BREAST CANCER DIGITAL REPOSITORY. In: 15th International Conference on Experimental Mechanics, Porto, Portugal, 2012.
48. Lotter W, Sorensen G, Cox D. A multi-scale CNN and curriculum learning strategy for mammogram classification. In: *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Cham: Springer; 2017. p. 169–77.
49. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, Back T, Chesus M, Corrado GC, Darzi A, et al. International evaluation of an ai system for breast cancer screening. *Nature.* 2020;577(7788):89–94.
50. Mikolov T, Deoras A, Povey D, Burget L, Cernocký J. Strategies for training large scale neural network language models. In: *Automatic speech recognition and understanding*. Waikoloa, HI, USA: IEEE; 2011. p. 196–201.
51. Moon WK, Huang YS, Hsu CH, Chien TYC, Chang JM, Lee SH, Huang CS, Chang RF. Computer-aided tumor detection in automated breast ultrasound using a 3-d convolutional neural network. *Comput Methods Prog Biomed.* 2020;190:105360.
52. Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS. INbreast: toward a full-field digital mammographic database. *Acad Radiol.* 2012;19(2):236–48.
53. Morrell S, Wojna Z, Khoo CS, Ourselin S, Iglesias JE. Large-scale mammography CAD with deformable conv-nets. In: *Image analysis for moving organ, breast, and thoracic images*. Cham: Springer; 2018. p. 64–72.
54. Oliver A, Freixenet J, Martí J, Perez E, Pont J, Denton ER, Zwiggelaar R. A review of automatic mass detection and segmentation in mammographic images. *Med Image Anal.* 2010;14(2):87–110.
55. Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *J Mach Learn Technol.* 2011;2:37–63.
56. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, IEEE, Las Vegas, NV, USA; 2016. p. 779–88.
57. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell.* 2017;39(6):1137–49. <https://doi.org/10.1109/tpami.2016.2577031>.
58. Ribli D, Horváth A, Unger Z, Pollner P, Csabai I. Detecting and classifying lesions in mammograms with deep learning. *Sci Rep.* 2018;8(1):4165.
59. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Teuwen J, Broeders M, Gennaro G, Clauser P, Helbich TH, Chevalier M, Mertelmeier T, Wallis MG, Andersson I, Zackrisson S, Sechopoulos I, Mann RM. Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. *Eur Radiol.* 2019;29:4825–32.
60. Shan J, Cheng H, Wang Y. Completely automated segmentation approach for breast ultrasound images using multiple-domain features. *Ultrasound Med Biol.* 2012;38(2):262–75.
61. Shayma'a AH, Sayed MS, Abdalla MI, Rashwan MA. Breast cancer masses classification using deep convolutional neural networks and transfer learning. *Multimed Tools Appl.* 2020;79(41):30735–68.
62. Shin HJ, Kim HH, Cha JH. Current status of automated breast ultrasonography. *Ultrasonography.* 2015;34(3):165.
63. Suckling J, Parker J, Dance D, Astley S, Hutt I, Boggis C, Ricketts I, Stamatakis E, Cerneaz N,

- Kok S, et al. The mammographic image analysis society digital mammogram database. In: International Congress series. Amsterdam: Excerpta Medica; 1994.
64. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proceedings of the conference on computer vision and pattern recognition. Boston, MA, USA: IEEE; 2015. p. 1–9.
65. Szegedy C, Toshev A, Erhan D. Deep neural networks for object detection. In: Advances in neural information processing systems. Lake Tahoe, USA: NIPS; 2013. p. 2553–61.
66. Tompson JJ, Jain A, LeCun Y, Bregler C. Joint training of a convolutional network and a graphical model for human pose estimation. In: Advances in neural information processing systems. Montréal CANADA: NIPS; 2014. p. 1799–807.
67. Vesal S, Patil SM, Ravikumar N, Maier AK. A multi-task framework for skin lesion detection and segmentation. In: OR 2.0 context-aware operating theaters, computer assisted robotic endoscopy, clinical image-based procedures, and skin image analysis. Cham: Springer; 2018. p. 285–93.
68. Yan Y, Sun X, Shen B. Contrast agents in dynamic contrast-enhanced magnetic resonance imaging. Oncotarget. 2017;8(26):43491.
69. Yap MH, Edirisinghe EA, Bez HE. A novel algorithm for initial lesion detection in ultrasound breast images. J Appl Clin Med Phys. 2008;9(4):181–99.
70. Yap MH, Goyal M, Osman F, Ahmad E, Martí R, Denton E, Juette A, Zwigelaar R. End-to-end breast ultrasound lesions recognition with a deep learning approach. In: Medical imaging 2018: biomedical applications in molecular, structural, and functional imaging, International society for optics and photonics, SPIE Medical Imaging, Houston, Texas, United States. vol. 10578; 2018. p. 1057819.
71. Yap MH, Goyal M, Osman F, Martí R, Denton E, Juette A, Zwigelaar R. Breast ultrasound region of interest detection and lesion localisation. Artif Intell Med. 2020;107:101880.
72. Yap MH, Goyal M, Osman FM, Martí R, Denton E, Juette A, Zwigelaar R. Breast ultrasound lesions recognition: end-to-end deep learning approaches. J Med Imaging. 2019;11007:1.
73. Yap MH, Pons G, Martí J, Ganau S, Sentís M, Zwigelaar R, Davison AK, Martí R. Automated breast ultrasound lesions detection using convolutional neural networks. IEEE J Biomed Health Inform. 2018;22(4):1218–26.
74. Zagoruyko S, Komodakis N. Wide residual networks. CoRR [abs/1605.07146](https://arxiv.org/abs/1605.07146). 2016. <http://arxiv.org/abs/1605.07146>
75. Zhou Y, Xu J, Liu Q, Li C, Liu Z, Wang M, Zheng H, Wang S. A radiomics approach with cnn for shear-wave elastography breast tumor classification. IEEE Trans Biomed Eng. 2018;65(9):1935–42.



Siva Teja Kakileti and Geetha Manjunath

Contents

Introduction	1302
Conventional Breast Imaging Techniques	1303
X-Ray Mammography	1303
Ultrasound	1303
Magnetic Resonance Imaging	1303
Challenges with Conventional Breast Imaging Modalities	1304
Infrared Thermography for Breast Imaging	1304
Breast Thermal Imaging Protocol	1305
Challenges with Manual Interpretation of Breast Thermography	1305
Artificial Intelligence for Breast Thermography	1307
AI for View Labeling	1307
AI for Breast Segmentation	1308
AI for Malignancy Classification	1308
AI for Risk Estimation	1310
AI for Biomarkers Prediction	1310
Discussion	1311
Conclusion	1313
References	1313

Abstract

The use of Artificial Intelligence (AI) in medicine has increased in the recent years due to its ability to aid the clinician for better clinical decision. AI solutions could play a critical role in the growing need for health care access and low clinician to population ratio. Breast cancer incidence rate is rising every year and there is an urgent need for scalable and low-cost solutions for breast cancer detection, especially for low

Supplementary Information: The online version of this chapter (https://doi.org/10.1007/978-3-030-64573-1_251) contains supplementary material, which is available to authorized users.

S. T. Kakileti (✉) · G. Manjunath
Niramai Health Analytix Private Limited, Bangalore, India
e-mail: sivateja@niramai.com; geetha@niramai.com

and middle income countries. Breast thermography has been used as an adjunct modality for breast cancer detection since 1980s. Breast thermography has advantages of being low-cost, non-invasive, radiation free, and privacy aware. However, subjectivity and high expertise required for interpretation of breast thermograms are major challenges for its poor adaptability and equivocal results in the literature for its effectiveness of breast cancer detection. To overcome these challenges, AI solutions have been proposed to automate different steps involved in breast thermography. The recent explorations have suggested that automated breast thermography can play a potential role in breast cancer detection, risk estimation, and hormonal status prediction. In this chapter, we discuss breast thermography and different AI applications in automated breast thermography.

Introduction

Cancer, a frequently used term for malignant neoplasms, can be described as uncontrolled growth of abnormal cells in a body. It occurs when normal cells in the body lose their specific functional behavior (*differentiation*), begin to multiply uncontrollably (*proliferation*), invade their host cellular matrix (*invasion*), and finally spread out into other adjacent or distant parts of the body (*metastasis*). Cancer is a medical condition that could permanently destroy the normal functioning of affected organs if not treated in the early stages. Breast cancer is a malignancy that develops in the cells of the breast (human mammary glands) tissue. It is typically formed either in the lobules that produce milk (lobular carcinoma) or in the milk ducts (ductal carcinoma) that carry milk towards the nipple [1].

Breast cancer is the leading cause of cancer deaths among women all over the world. According to the World Health Organization (WHO), nearly 627,000 women lost their lives to this disease in 2018 [2, 3]. The incidence rate has been placed at 2.1 million every year [2], thus averaging to approximately one woman being diagnosed with breast cancer out of every 17 women in her lifetime. It is estimated that the incidence and mortality of breast cancer could

increase to 3.06 million and 0.99 million, respectively, by 2040 [4]. These incidence and mortality rates further vary with countries and their economic status [2, 3]. Breast cancer can be caused through genetic mutations that could be hereditary or acquired during the lifetime, either randomly or induced by exposure to carcinogens such as certain industrial chemicals and excessive ionizing radiation like X-rays [1]. As summarized by Feng et al. [1], obesity, hormonal replacement therapy, use of contraceptives for birth control, alcohol consumption, early menarche, and late childbirth are other significant risk factors for breast cancer [1]. It has approximately a hundred times higher incidence among women compared to men, and the risk of breast cancer increases with age [1].

It has been shown in various research studies that early detection of breast cancer can improve the chances of survival significantly [5]. The conventional imaging modalities for breast cancer detection include mammography, ultrasound, and Magnetic Resonance Imaging (MRI) and are discussed in detail in section “[Conventional Breast Imaging Techniques](#).” However, high cost equipment, need for high skilled personnel, dependence on infrastructure, use of ionizing radiation, among others, are major hurdles for their implementation, as discussed in section “[Challenges with Conventional Breast Imaging Modalities](#).” Expected increase in the incidence of breast cancers in the coming years [4] calls for acute need for early breast cancer detection techniques that are automated, affordable, and portable to improve accessibility.

Infrared breast thermography, which measures breast surface temperatures, has been used for breast cancer detection as an adjunct modality (section “[Infrared Thermography for Breast Imaging](#)”). Breast thermography can address some of the potential challenges faced by the conventional imaging modalities. However, subjectivity and high expertise required for interpretation of breast thermograms are major challenges for its poor adaptability. With the advances in Artificial Intelligence (AI) and improved infrared detectors in the last two decades, breast thermography is emerging as a mainstream imaging modality for breast cancer detection. In this chapter, we take a critical look at manual thermography (section “[Challenges with Manual Interpretation of Breast Thermography](#)”) and recent

advancements of automated thermography with Artificial Intelligence (section “[Artificial Intelligence for Breast Thermography](#)”) and discuss the potential use of automated breast thermography for breast cancer detection (section “[Discussion](#)”).

Conventional Breast Imaging Techniques

X-Ray Mammography

Mammography is considered as the gold standard technique for breast cancer detection. It uses planar projection X-ray images to visualize the internal structure of the breast. The cancerous cells generally are high-density regions and appear as white or enhanced regions in the mammographic images. Mammography uses low doses of X-rays to study two main types of breast changes such as high-density regions and calcifications. There have been many randomized trials in the literature that show the effectiveness of mammography for breast cancer detection [5]. There are also various studies that report the role of mammography in reducing the mortality of breast cancer [5, 6]. In general, the sensitivity and specificity of mammography reported in these studies for breast cancer detection vary from 64% to 90% and 82% to 98%, respectively, across different populations [6–8].

Unfortunately, mammography has some disadvantages. The use of ionizing radiation for breast imaging can increase the risk of secondary cancer. This led to the restriction of repeated usage of mammography, and the United States Preventive Services Task Force (USPSTF) and the European Commission have regulated the usage of X-ray mammography to once in every two years (on an average) across different age groups. According to some studies [8–10], mammography has sensitivity less than 70% and a high false positive rate of 56% in young women who generally have high density breast tissues. Since the breast tissue surrounding the lesion is also highly dense, it would be significantly harder to detect the cancerous lesion in these women. Also, to image the entire breast region, 15 to 20 pounds of pressure is applied onto a breast to flatten it against an imaging receptacle, which can be very uncomfortable and often painful for women.

A survey by Forbes et al. [11] stated that physical pain, embarrassment, inconvenience, and physical discomfort involved during breast compression are among the reasons why women may defer mammography or prefer to avoid mammography.

Ultrasound

Ultrasound is a commonly used breast imaging modality to characterize a mass or an abnormality inside the breast region. It uses sound waves well above the frequency range of human hearing to investigate the structural properties of soft tissue such as muscle and fat. The invasive nature of cancer cells to spread to the surrounding host tissue makes their structure particularly unusual compared to smooth oval and circular structures expected of healthy breast tissues. One can also observe the microscopulations on the host-tumor boundary region indicative of tumor aggressiveness [28]. The sensitivity and specificity of standalone ultrasound from different studies for breast cancer detection varies from 53% to 75% and 89% to 98%, respectively [7, 8]. The new advances of ultrasound imaging are further aiding in improving the accuracy of ultrasound imaging.

In clinical practice, ultrasound imaging is typically used for corroborative findings alongside mammography and/or clinical examination by manual palpation of the breasts. It is found to increase sensitivity by more than 20% in detecting small cancers and tumors in dense breasts when used in conjunction with mammography than mammography alone [7, 8]. However, high technical and clinical expertise is necessary to interpret the ultrasound images [8, 12], along with a dedicated mean examination time of 10 to 20 min from a clinician [12].

Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) captures detailed anatomical information of the body using strong magnetic fields and nonionizing radio waves. The nuclei of hydrogen atoms in the body emit very weak radio waves when excited in a strong magnetic field. The frequency of these emitted radio waves is indicative of the immediate

physico-chemical environment of the water molecules. Radio wave pulses from an external transducer are used to synchronize and trigger the emission of the nuclear radio waves, such that an image can be progressively reconstructed from the frequency of the emitted radio waves.

Several studies showed that breast MRI can result in very high sensitivity (over 90%) for detecting cancerous abnormalities [12]. However, it has high false positive rates with low to moderate specificity ranging from 37% to 72% [13]. This could also be the reason why MRI is not used as a routine screening test for breast cancer detection in an otherwise asymptomatic or low risk population. In general, women with particularly higher risk of breast cancer and high breast density are recommended for MRI examination along with mammography to monitor the variations over time in the breast tissue [13]. However, MRI scanners are expensive and require dedicated infrastructure and skilled staff for their usage. Furthermore, lying still for extended periods inside the physically constrained environment of the MRI scanner imaging bore may not be suitable for body habitus or may cause claustrophobia.

Challenges with Conventional Breast Imaging Modalities

The diagnostic and screening implementations with the above conventional imaging modalities have demonstrated a reduction in mortality rates due to breast cancer [5, 6, 8], especially in high income countries. However, there are some challenges with their implementation especially in low- and middle-income countries. Some prominent reasons [6, 8] being:

1. *High equipment cost and high maintenance cost* of conventional screening and diagnostic facilities limits their availability to very few specialist clinics in developing countries. Non-portability of the equipment and the dependence on infrastructure further limits their usage to conduct screening camps.
2. Need for *high skilled* technicians to operate machines such as mammography, ultrasound, and MRI and for *experienced radiologists* to

interpret the imaging data generated. In India, for example, there are less than 10,000 radiologists for a population of 1.3 billion, that is, a single radiologist per 100,000 people [14].

3. Restrictions on usage of mammography and ultrasound. Mammography is not recommended for frequent imaging due to the use of ionizing radiation. Ultrasound is restricted in some countries since it can be used for fetal sex determination.
4. Lack of *awareness* among women to go for breast cancer screening.
5. *Privacy or embarrassment* during the breast imaging is found to be one of the major reasons why women, especially from South and South East Asian countries, do not prefer to attend routine breast cancer screening [11]. A survey conducted by Forbes et al. [11] found that 59% of Indian women reported embarrassment as a barrier for screening.
6. *Social or cultural barriers* such as stigma, fear, and cancer fatalism [6].

Infrared Thermography for Breast Imaging

Breast infrared thermography is considered as a low-cost imaging modality that has found its usage for breast cancer detection since the 1960s [15]. The high metabolism required for the growth of cancer cells is promoted by the increased resources to the vicinity of the tumor region. To facilitate the high resource consumption, new blood vessels are formed, large volumes of blood is circulated through the existing blood vessels, and dormant blood vessels are recruited [15]. The increased metabolic activities generate more heat and the generated heat is transferred to the breast surface through tissue conduction and venous convection. The high temperatures associated with cancerous regions and blood vessel regions were validated experimentally in various studies [16]. Therefore, it might be possible to detect the cancerous lesions by investigating the surface breast thermal patterns. Breast thermography allows a noninvasive and noncontact way of breast imaging which further enables a privacy aware imaging. Further, thermal imaging

equipment is affordable and portable, addressing the infrastructure and affordability issues with the conventional screening modalities such as mammography and MRI.

In the literature, there have been various studies that show the effectiveness of breast thermography conducted under controlled settings for breast cancer detection [15]. In different independent studies [15–17], it was also observed that thermography was able to identify cancerous lesions before they show positive signs in conventional imaging. As the metabolic changes occur from the onset of cancer, it might allow thermography to detect cancers in early stages even before a mass is formed. Based on this hypothesis, some studies [15–17] validated if thermography can be used as a risk indicator to identify women who are at high risk. Anbar et al. [18] found that dynamic thermography, videos at high frame rate, can be used to differentiate cancer types such as invasive, *in situ*, and inflammatory cancers. In a recent exploration, Zore et al. [19] observed that thermal imaging shows different heat patterns for hormonal positive and negative tumors, a biomarker used for prognosis, and treatment planning. This could be due to the dependence of hormones on the increased heat response.

Appreciating these potential implications in breast imaging, the Food and Drug Administration (FDA) approved the usage of thermography as an adjunctive diagnostic imaging modality for breast cancer detection in 1982. The significant advances in infrared detectors involving uncooled microbolometer Focal Plane Arrays (FPA) over the last two decades have led to the development of low-cost infrared imaging devices that can detect minute temperature variations up to 20 mK. This has further aided in improving the performance of breast thermography in breast imaging as discussed in [15, 20].

Breast Thermal Imaging Protocol

As summarized in [15], imaging protocols for breast thermography can be broadly categorized into continuous and discrete imaging protocols. Continuous imaging protocols involve analyzing the thermal signatures over time by capturing

videos at high frame rate. On the other hand, discrete imaging protocols involve analyzing a set of views captured at different view angles. In the literature, discrete imaging protocol is prominently used for breast thermography. Typically, five views at 0° (Frontal), $+45^\circ$ (Left Oblique), $+90^\circ$ (Left Lateral), -45° (Right Oblique), and -90° (Right Lateral) are captured to image the 180° of the breast region as shown in Fig. 1.

Challenges with Manual Interpretation of Breast Thermography

As discussed in the previous section, breast thermography can be used to detect breast abnormalities. However, interpretation of thermograms requires high expertise and is therefore subjective. Since thermography captures the temperatures emitted from a surface, the captured thermal data would comprise of real-valued temperature measurements with a precision dependent upon the thermal sensitivity of the infrared device. Therefore, pseudo color palettes like Rainbow, Iron, Gray, etc., are used for visual interpretation [21]. A lower and upper temperature limits need to be set to convert the captured thermal data into a pseudo color image for visualization. Any change in the temperature limits could change the pseudo color image, as illustrated in Fig. 2, and hence, might affect the interpretation of thermal images. This could introduce subjectivity in interpretation and the visual inspection of different color shades representing minute temperature variations could be a high cognitive load for the clinician. Further, it is not trivial to classify malignancy by observing the color shades of the pixels with naked eye though thermal abnormalities appear as hot or high intensity regions. This is due to evidences that some of the benign conditions like duct ectasia, fibroadenoma, fibrocystic, abscess, and high hormonal response could cause high metabolic heat generation and appear as high intensity regions on breast thermal images [22]. These factors might be a reason for the existence of contradicting results about the use of thermography for breast cancer detection.

The Breast Cancer Detection Demonstration Project (BCDDP) [23], a breast cancer screening

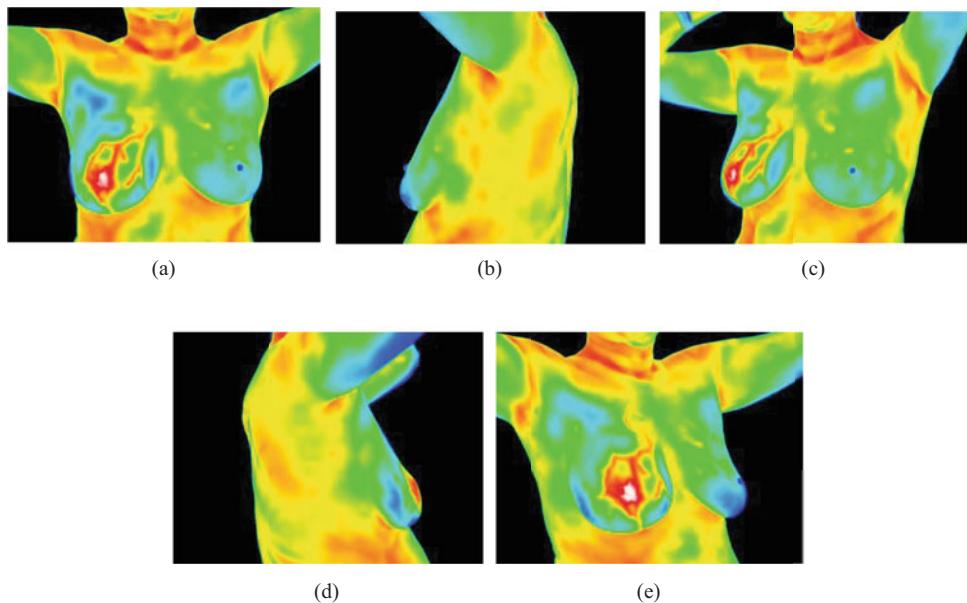


Fig. 1 Thermal images of a subject in Meditherm color palette captured at (a) frontal, (b) left lateral, (c) left oblique, (d) right lateral, and (e) right oblique views

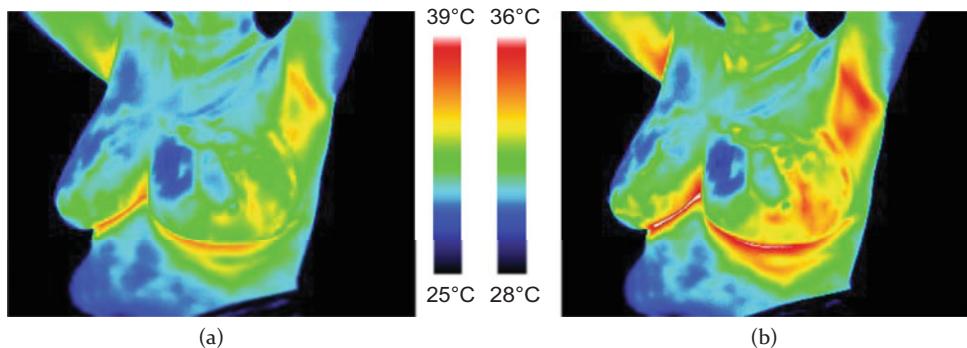


Fig. 2 Thermal images of a participant with different temperature ranges. (a) 25 °C to 39 °C. and (b) 28 °C to 36 °C

program conducted between 1973 and 1980, and few more studies [24] reported poor performance of standalone thermography for breast cancer detection. According to Gautherie et al. [16], Keyserlingk et al. [25], Kennedy et al. [26], and Gonzalez et al. [20], these poor results could be due to the lack of adequate quality control, technical skills, training for using infrared equipment and high expertise required to interpret breast thermograms. According to Keyserlingk et al. [25], the low sensitivity could also be attributed to the failure in the image capture of oblique views, which could have led to missed cancers

in the outer quadrants of breasts. As discussed by Gonzalez et al. [20], some of the low accuracies reported in the studies could be due to low sensitive cameras used for the study and the accuracies could have been improved with the latest advances in the infrared thermal cameras that can measure minute variations up to 20 mK. The detection of minute variation in temperatures would allow for better visualization and discrimination of vascular and nonvascular heat patterns and could improve thermographic category grading [25]. Further, most of the studies that reported low sensitivities were performed during 1970 and

1990 [20]. Overall, from these observations, the accuracy or resolution of thermography seems to be benefited from the advances in infrared detectors; however, the accuracy obtained with thermography seems to be subjective and requires high expertise for interpretation of breast thermal images. This may be a key reason for its poor adoption in the clinical practice.

These reasons have led the Food and Drug Administration (FDA) to impose restrictions and stringent warnings [27] on the use of breast thermography as a standalone imaging modality for breast cancer detection. In order to reduce these false thermal interpretations and subjectivity, we need a robust computer aided system that can learn expert domain knowledge and help in automating the interpretation of breast thermal images. The high resolution data from recent highly sensitive thermal cameras help the computer-aided algorithms in analyzing minute heat variations that might be difficult to interpret with the naked eye.

Artificial Intelligence for Breast Thermography

Artificial Intelligence (AI), the ability of computer systems to perform tasks that typically require human intelligence, has emerged significantly over the recent years and has started playing a pivotal role in various sectors. Specifically, there has been a tremendous increase in the investigations of AI applications for healthcare [28] in the last decade. Though there were many speculations regarding the implications of AI on patient care in the beginning, the recent regulatory clearances for AI based medical devices from regulatory boards such as the US FDA indicate its strong role in the future of healthcare [29, 30]. In breast radiology, there have been several AI algorithms in the literature for automated analysis of conventional imaging to reduce subjectivity and improve accuracy. Some of these AI applications such as Quantx [31] and Transpara [32] are cleared by FDA for analyzing MRI and mammography images, respectively. Overall, the use of AI in radiology is receiving appreciation and seems to alleviate the need for high expertise required to interpret the images and

helps in reducing the overall time required by a radiologist for examination.

Breast thermography has many challenges starting from image capture to interpretation of thermal images. Like many other imaging modalities, breast thermography captures multiple thermal images at different view angles, as described in section “[Infrared Thermography for Breast Imaging](#).” This could lead to subjectivity in thermal imaging capture and mislabeling of captured views. As discussed in the previous section, manual interpretation of thermograms is difficult and subjective and might be a reason for the controversial results observed with manual analysis of breast thermography. Therefore, AI can help in addressing the interpretation issues with manual thermography and might help in automating, scaling and standardizing the thermal image interpretation for breast cancer detection.

Subjectivity with manual thermography, availability of low-cost infrared camera equipment coupled with the recent advances in the infrared technology, and the radiation-free benefits of thermography have led to the sparked interest in the machine learning community on automated breast thermography from the 2000s. As summarized in recent review papers [33, 34], and a PhD Thesis [63], there are various studies in the literature that show the effectiveness of AI in various phases of automated interpretation of breast thermal images. These AI applications for thermography can be categorized into five key tasks, which are detailed in the following subsections.

AI for View Labeling

Image capture plays a crucial role in obtaining accurate results with any computer-aided diagnosis/detection solutions. As mentioned in section “[Breast Thermal Imaging Protocol](#),” breast thermography involves the capture of multiple images of the subject at five different view angles. Thermographer or technician needs to manually save these captured images corresponding to their view angles to enable automated analysis. Manual saving of the images could be erroneous due to the presence of the right/left breast on the left/right

side of the image and might result in inaccurate analysis. To deskill this process, there has been a recent approach [35] that uses ResNet 50, a 50-layer convolutional neural network (CNN) with residual skip connections, to automatically predict the view from the thermal image. The results of 97% accuracy for 5-class classification corresponding to the five captured views and 99.5% accuracy for labeling into left, right, and frontal views show the potential of AI in automated view labeling of thermal images. This automation can help in decreasing the overall time for image capture and also to flag any errors in cases of wrong or improper view capture.

AI for Breast Segmentation

Breast segmentation is essential for automated analysis of thermal patterns inside the breast region. The manual segmentation of a breast region is laborious and is subjective due to the lack of strict breast boundaries. The minimal variation in thermal intensities between the breast and the nonbreast regions and amorphous nature of the breast make the automated segmentation of thermal images a challenging problem. In the existing literature, there have been many automated and semi-automated approaches for breast segmentation [33–37]. These techniques typically use traditional image processing approaches to segment the breast regions by using heuristics from inframammary fold region, axilla region, and

breast silhouette boundary [33–37]. The sensitivity, accuracy, and dice index reported in the literature were greater than 90%. Most of these techniques work either for a frontal or a lateral view. A recent approach shows that an encoder-decoder CNN can be used to segment the breast region in multiple views [38]. Specifically, authors propose the use of V-Net and showed that a Dice index of 0.92 can be achieved for segmentation of breast region. They further explain that the obtained Dice index is 1% higher than the interobserver dice correlation. This shows that AI has potential to perform similar to a human expert. To further refine the AI predicted segmentation with human intelligence, some authors suggest an elliptical or polygon user interface on the AI output [38] for manual adjustment, as shown in Fig. 3. Overall, these advances in AI are transforming the laborious and time-consuming breast region segmentation into a simple and fast process.

AI for Malignancy Classification

Malignancy classification involves classification of breast thermal patterns into two classes: malignant or benign. As discussed in section “[Infrared Thermography for Breast Imaging](#),” breast heat patterns are affected by the increased vascularity and metabolic activity of the cancerous region. However, the increase in the heat patterns can be observed with some of the benign conditions as well. This makes the manual interpretation

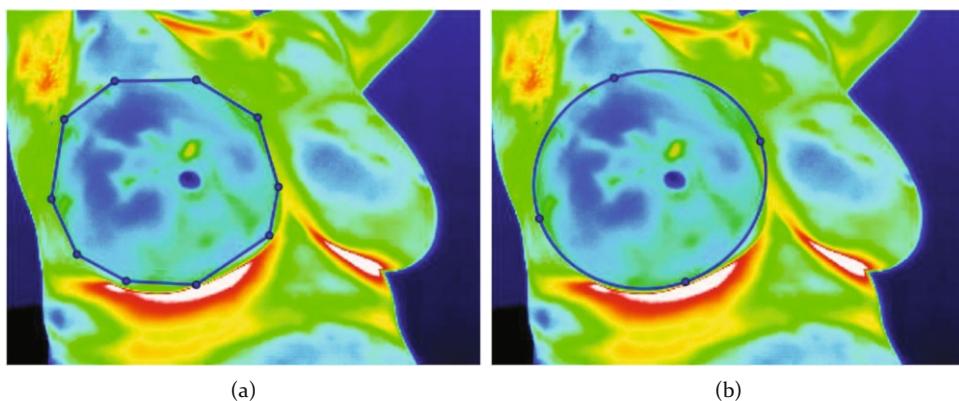


Fig. 3 Manual adjustment of the predicted breast region with (a) polygon, and (b) ellipse

challenging and error prone. Automated classification of malignancy from the heat patterns could help in alleviating the subjectivity and reducing the false interpretations that were earlier reported with manual thermography. The problem of differentiating the malignant and benign heat patterns with AI is well explored in the literature and there are many AI approaches to classify the heat patterns. The sensitivities and specificities reported in the literature vary from 79% to 100% and 79% to 100%, respectively [20, 33, 34], showing the feasibility of automated interpretation. These AI solutions differ in the features used for classification and broadly can be divided as below:

Textural and statistical features: These methods extract textural and statistical features describing the overall heat patterns on the breast region and classify using Machine Learning (ML) classifiers such as Support Vector Machines (SVM), Random Forests (RF), Artificial Neural Networks (ANN), Multi-Layer Perceptron (MLP), etc. These features involve mean, contrast, moments, energy, homogeneity, entropy, etc., of the heat patterns inside the breast region [33, 34]. Some approaches use Histogram of Oriented Gradients (HOG) features of the breast region for classification [39] into benign and malignant.

Hotspot features: These techniques first identify high thermal regions (hotspots) and extract features from these regions to characterize malignant heat patterns. Clustering techniques such as

fuzzy c-means and k-means [40], active Contours [41], adaptive histogram thresholding [22], and CNNs [42] have been explored in the literature to automatically segment the hotspot regions. From these hotspots, domain knowledge features [22, 43] such as irregularity, symmetry, temperature increase, etc., for differentiating between benign and malignancy are proposed for classification using ML classifiers. The detection of hotspots helps in localization of tumor region, whereas extraction of domain knowledge features from hotspots might improve semantic interpretation by a clinician. Figure 4a shows the high and warm thermal regions detected by the approach proposed by Madhu et al. [22] for a sample thermal image.

Vascular features: Vascular deformations are the initial changes that are associated with the abnormal growth of cancer cells [15]. Thermal imaging can be used to study these vascular deformations. The heat generated from the blood vessel activity might be diffused during its transmission to the breast surface, and therefore, the captured heat pattern might have diffused vessel boundaries. Therefore, automated vessel extraction techniques in other imaging domains might not work well for breast thermal images due to these diffused vessel boundaries. Due to this reason, authors in [44–46] propose multiple structural and intensity enhancement techniques to identify the blood vessel structures, as illustrated in Fig. 4b. From the detected vessels, authors [44, 45] suggest the use of domain features such

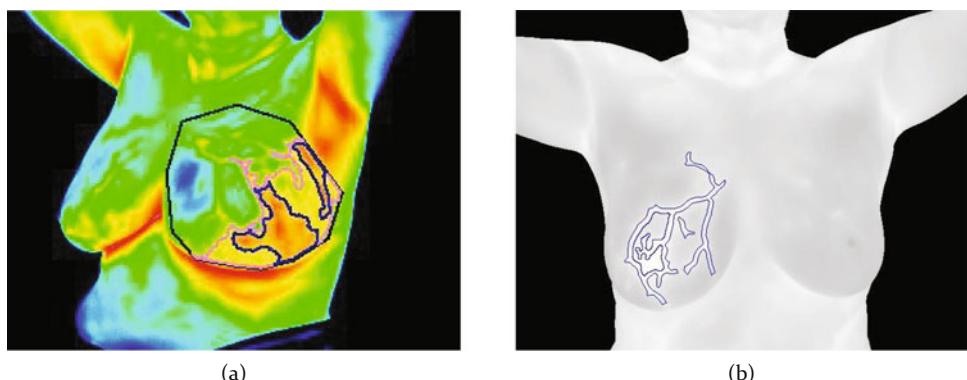


Fig. 4 (a) High and warm thermal regions detected by the approach proposed in [22]. (b) Vessel like structures detected by the approach proposed in [44]

as tortuosity, symmetry, number of blood vessels, etc., for classification of malignancy with ML classifiers. In [46], authors compare thermal minutiae points of detected vessels before and after the application of cold challenge for classification of malignancy.

Pixel level classification: This class of algorithms classifies each pixel or block of pixels in an image as malignant or normal. In [47], authors use statistical features such as signal complexity and signal mobility of the heat variations at each breast pixel on the image over time. In [48], authors propose a split and merge approach using a multifeature fusion algorithm to detect the tumor automatically. Since each pixel or block of pixels is being classified, these set of algorithms can be used for localization of malignancy.

CNN-based features: CNNs eliminate the need for hand-crafted features by learning the representation between the input images and their corresponding output labels. In the literature, authors [49–52] have used deep learning architectures such as ResNet, VGG, Inception, GoogLeNet, DenseNet, MobileNet, and ShuffleNet to predict the malignancy from thermal images. These deep learning architectures are showing promising results for prediction of malignancy from thermal images. However, interpretation of the predicted outcome is a known challenge with deep learning. In a very recent exploration [52], authors propose the use of activation maps to identify focus areas of deep learning architecture for predicting malignancy. However, the validation of focus areas is not performed. Overall, these results are encouraging, and significant research is being conducted to further improve the interpretation with deep learning.

AI for Risk Estimation

Breast cancer risk estimation can help in identifying women with high risk of breast cancer and thereby aiding the clinician to create a personalized care for women tailored to their level of risk. The identification of high-risk women can be extremely beneficial in creating awareness

among women and to divert the limited resources for those who would need them the most in rural camps. As breast thermography captures the metabolic information in the form of heat patterns in the breast region of the subject, it might be used for breast cancer risk estimation. A clinical study conducted by Gautherie et al. [16] with manual interpretation reported that approximately one-third of 1,245 women who had an equivocal thermogram and a normal examination by conventional means in their first visit had developed breast cancer in the next five years. In the literature, there are both automated and semi-automated AI approaches to evaluate the use of thermography for breast cancer risk estimation. In [53], authors used a semi-automated assessment criterion called BIRAS (Breast Infrared Assessment System) to categorize the thermal images into five cohorts with BIRAS 1 being low risk and BIRAS 5 being high risk. This approach considers 10 images (5 before cooling and 5 after cooling) and uses automated nonvascular criteria such as symmetry, areolar temperature, isothermia in each quadrant, temperature decrease due to cooling and hotspot parameter and vascular criteria described by the clinician. In a recent exploration [45], a completely automated risk estimation called Thermalytix Risk Estimation (TRS) is proposed using vascular features, non-vascular features, age, and symptoms for pre-screening of breast cancer. Authors suggest that the estimated TRS can then be used to categorize women into four risk cohorts. In both these approaches, the proportion of malignancies has increased with increased risk showing the potential role of automated breast thermography for risk estimation.

AI for Biomarkers Prediction

Biomarkers involving estrogen (ER), progesterone (PR), KI67, and HER2 play an important role in the prediction of prognosis and treatment procedure for breast cancer. Typically, histopathology of tissue samples that are extracted through a surgical procedure is used for assessing the biomarkers. Obtaining these biomarkers

would be costly and time consuming for the patient due to the limited pathology labs in the developing countries. In a study conducted by Zore et al. [19], it is observed that tumors with positive hormonal (estrogen/progesterone) status have a different thermal signature compared to hormonal negative tumors. To automate the hormonal status classification with AI, authors [54] propose the use of machine learning features such as symmetry, Jensen Shannon Divergence (JSD), relative hotness, thermal distribution ratio, and textural features. This automation can make biomarkers estimation an automated, fast, and non-invasive approach.

Discussion

The advanced machine learning solutions, as discussed in the previous section, are aiming to generate an increasingly consistent and accurate interpretation of breast thermograms. The use of AI in medical imaging can allow continuous learning through retraining of the machine learning models from a new variety of data. This could be essential to accustom the AI solution to different cancer types and to the changes in the cancer mutations over time. The improved accuracies in the literature for automating breast thermography with the advancements of AI over the recent years are showing a promising role of breast thermography in breast cancer care.

The lack of publicly available large datasets with curated ground truth seems to be a major challenge for research in automated breast thermography. The availability of PROENG [55] dataset helped the machine learning community to experiment on breast thermography for segmentation and classification. However, the small training dataset has limited most of the existing approaches to use traditional image processing and machine learning for different problems in breast thermography. In the presence of large datasets [35] and with the techniques of transfer learning of weights from one domain to other domain [49–52], a transition has seen in recent years from traditional image processing and machine learning approach to deep learning.

While the former might require a small amount of data, it might not be generalizable. On the other hand, deep learning can produce better results but they might overfit to the input distribution in presence of smaller training datasets. This brings a trade-off between these approaches and evaluating of these techniques would be critical in measuring the real effectiveness. In general, evaluation of results on independent test sets rather than the validation datasets is commonly used to compare the approaches. The ground truth for training and validating the ML models is another important factor in measuring the effectiveness of these ML classifiers. In many studies, the ground truth for classification is not described in detail. Since thermography is not a standalone modality, comparison with manual labeling would not result in actual sensitivity and specificity of automated breast thermography for breast cancer detection. Biopsy is considered to be the ultimate ground truth for breast cancer detection. However, it is not ethical to perform biopsy on all the subjects without any strong suspicion from malignancy. To tackle this, a combination of conventional imaging modalities such as mammography, ultrasound, and MRI can be used to confirm the negative ground truth, whereas biopsy can be used to confirm the positive ground truth.

Most of the existing AI approaches for different tasks in breast thermography considered breast thermal images of women who had no prior history of breast cancer. This might limit the applicability of these approaches on women who had surgical procedures like mastectomy, breast augmentation, breast reduction, lumpectomy, etc. Segmentation and tagging that rely on boundary features might not work on these women as the breast contours would be significantly modified. Also, the heat patterns might be disturbed due to surgery and might interfere with the classification of heat patterns into benign or malignant. Screening of these women with prior breast cancer is important due to the chances of recurrence of breast cancer either in the same or contralateral breasts. As mammography might not be recommended for mastectomized or recently lumpectomized women, it would be interesting to see if breast thermography can be effective in

these women. Similarly, the ML approaches should also consider male population distribution during training of the ML models as 1 in 1000 men have a chance of breast cancer.

Further, the adaptation of automated thermography in clinical practice is still in its initial stages. Currently, most of the evaluations of AI solutions for breast thermography are done retrospectively and there are only a few studies [56–59] that have evaluated the automated AI solutions in clinical settings. In a recent study [60], authors evaluated the effectiveness of an AI tool called Thermalytix on symptomatic asymptomatic cohorts at multiple sites, where they reported a sensitivity of 91% for detecting breast cancers. However, most of these existing studies involved small cohorts with similar demography. Therefore, it is important to validate breast thermography AI solutions on large-scale diversified populations for their clinical effectiveness in various stages of breast cancer care continuum before its clinical usage. Specifically, the role of breast thermography in screening, diagnosis, and prescreening needs to be evaluated through randomized clinical trials. Though the cost of breast thermography equipment is low compared to conventional screening, specificity ($1 - \text{false positive rate}$) of breast thermography with AI solution is found to be low in some studies. A detailed health technology

assessment needs to be conducted to understand the real cost effectiveness of breast thermography.

Interpretability of a predicted outcome is another crucial factor for clinicians as it could play a critical role in patient care [61, 62]. There has not been any literature that talks about the interpretability and user/clinician studies for the existing automated breast thermography systems. As observed in a study conducted by Bratko et al. [61], clinicians do not want to rely solely on the predicted outcome but would be interested in simplified and sensible qualitative interpretation. According to Vellido et al. [62], visualization can aid a perfect conduit for the interpretation of algorithmic data. In addition, a cyclic feedback through an interactive visualizer could aid in dealing with detected outliers, anomalous data, or data artifacts. From these studies, it is evident that a computer-aided thermographic system that could estimate quantified qualitative thermal parameters with a supported visual interpretation in addition with a cyclic feedback is necessary for better and quick clinical acceptance. Figure 5 shows a possible cyclic feedback for integrating different modules of breast thermography for better clinician interaction. It is essential to experiment and study different cyclic integrations of automated thermography systems for its implications on clinical adaption and improving the overall accuracy.

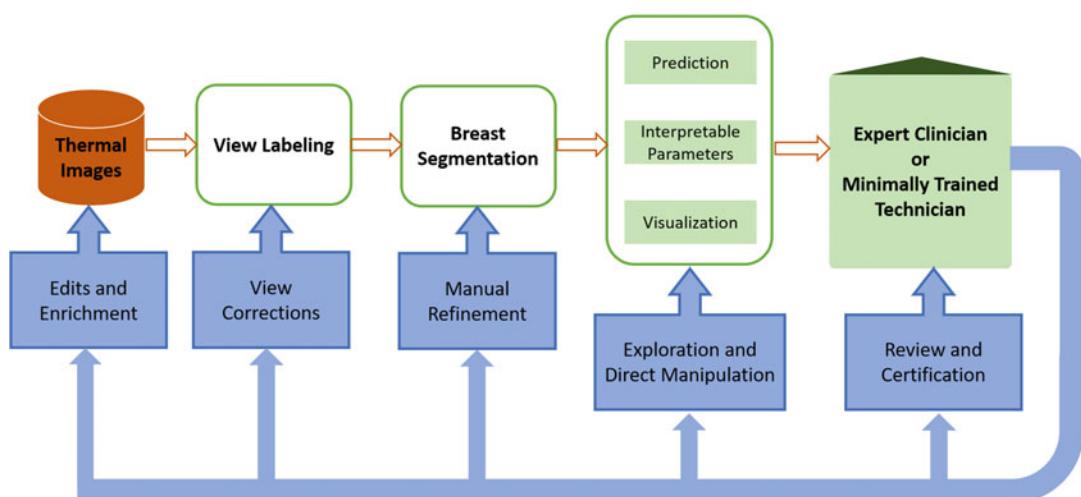


Fig. 5 A block diagram showing different AI modules with cyclic feedback in automated breast thermography for improving clinical interpretation

This need for interpretation also questions the usage of deep learning as it might be difficult to explain the predicted outcome. Though the activation maps, as discussed in [52], are improving the interpretability of deep learning, proper validation needs to be done for their accuracy and implications on interpretation.

In addition to the potential applications of AI for breast thermography as discussed in section “[Artificial Intelligence for Breast Thermography](#),” breast thermography was used for treatment monitoring [25], bio-heat simulation [20], and cancer subtype determination [18]. However, the results are preliminary and there have not been any clinical studies validating these applications. Further, transfer learning can be explored to check if ML classifiers trained with breast thermography data can be used to validate the feasibility of automated thermography for the detection of other cancers such as oral cancer, thyroid cancer, and skin cancer. At last, there is still a long road ahead for automated thermography; much improvement, validation, and large-scale clinical studies are needed before its implementation into clinical practice.

Conclusion

Breast infrared thermography has benefits of low-cost, radiation-free, noninvasiveness, privacy awareness, and portability. These advantages could become a boon especially to low- and middle-income countries. Subjectivity and potential misinterpretation with breast thermography have led to stringent restrictions on its usage and poor adaptability in the medical community. However, the advancements of infrared detectors when combined with the recent developments in Computer Aided Diagnostics (CAD) on thermal images show promising improvements in generating consistent and accurate results. There are a few systematic clinical studies that show the effectiveness of such an end-to-end automated CAD system for breast cancer detection. For clinical adaptation of such automated tools for assisted interpretability, it is important to evaluate these CAD systems on a large-scale diverse population.

Acknowledgements We would like to thank Prof. Andre Dekker and Prof. Leonard Wee from Maastricht University, Netherlands, for their valuable suggestions and comments on the work.

References

1. Feng Y, Spezia M, Huang S, Yuan C, Zeng Z, Zhang L, Ji X, et al. Breast cancer development and progression: risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. *Genes Dis.* 2018;5(2):77–106.
2. <https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/>
3. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2018;68(6):394–424.
4. Cancer Tomorrow. GLOBOCAN 2018. Retrieved from: <http://gco.iarc.fr/>. Accessed 28 June 2020.
5. World Health Organization. WHO position paper on mammography screening. Geneva: World Health Organization; 2014.
6. Coleman C. Early detection and screening for breast cancer. In: Seminars in oncology nursing, vol. 33(2). United Kingdom: WB Saunders; 2017. p. 141–55.
7. Kolb TM, Lichy J, Newhouse JH. Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: an analysis of 27,825 patient evaluations. *Radiology.* 2002;225(1):165–75.
8. Rajaraman P, Anderson BO, Basu P, Belinson JL, D'Cruz A, Dhillon PK, Gupta P, et al. Recommendations for screening and early detection of common cancers in India. *Lancet Oncol.* 2015;16(7):e352–61.
9. Carney PA, Miglioretti DL, Yankaskas BC, Kerlikowske K, Rosenberg R, Rutter CM, Geller BM, et al. Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. *Ann Intern Med.* 2003;138(3):168–75.
10. Yankaskas BC, Haneuse S, Kapp JM, Kerlikowske K, Geller B, Buist DSM, Breast Cancer Surveillance Consortium. Performance of first mammography examination in women younger than 40 years. *JNCI J Natl Cancer Inst.* 2010;102(10):692–701.
11. Forbes LJL, Atkins L, Thurnham A, Layburn J, Haste F, Ramirez AJ. Breast cancer awareness and barriers to symptomatic presentation among women from different ethnic groups in East London. *Br J Cancer.* 2011;105(10):1474–9.
12. Lee CH, Dershaw DD, Kopans D, Evans P, Monsees B, Monticciolo D, Brenner RJ, et al. Breast cancer screening with imaging: recommendations from the Society of Breast Imaging and the ACR on the use of mammography, breast MRI, breast ultrasound, and other

- technologies for the detection of clinically occult breast cancer. *J Am Coll Radiol.* 2010;7(1):18–27.
13. Menezes GLG, Knutel FM, Stehouwer BL, Pijnappel RM, van den Bosch MAAJ. Magnetic resonance imaging in breast cancer: a literature review and future perspectives. *World J Clin Oncol.* 2014;5(2):61.
 14. Arora R. The training and practice of radiology in India: current trends. *Quant Imaging Med Surg.* 2014;4(6):449–50.
 15. Kakileti ST, Manjunath G, Madhu H, Ramprakash HV. Advances in breast thermography. In: *Breast imaging: new perspectives in*. Rijeka: IntechOpen; 2017. p. 91.
 16. Gautherie M, Gros CM. Breast thermography and cancer risk prediction. *Cancer.* 1980;45(1):51–6.
 17. Etehadtavakol M, Ng EYK. Breast thermography as a potential non-contact method in the early detection of cancer: a review. *J Mech Med Biol.* 2013;13(02):1330001.
 18. Anbar M, Milesu L, Naumov A, Brown C, Button T, Carly C, AlDulaimi K. Detection of cancerous breasts by dynamic area telemeteretry. *IEEE Eng Med Biol Mag.* 2001;20(5):80–91.
 19. Zore Z, Boras I, Stanec M, Orešić T, Zore IF. Influence of hormonal status on thermography findings in breast cancer. *Acta Clin Croat.* 2013;52(1):35–42.
 20. Gonzalez-Hernandez J-L, Recinella AN, Kandlikar SG, Dabydeen D, Medeiros L, Phatak P. Technology, application and potential of dynamic breast thermography for the detection of breast cancer. *Int J Heat Mass Transf.* 2019;131:558–73.
 21. Lahiri BB, Bagavathiappan S, Jayakumar T, Philip J. Medical applications of infrared thermography: a review. *Infrared Phys Technol.* 2012;55(4):221–35.
 22. Madhu H, Kakileti ST, Venkataramani K, Jabbireddy S. Extraction of medically interpretable features for classification of malignancy in breast thermography. In: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE; 2016. p. 1062–5.
 23. Baker LH. Breast cancer detection demonstration project: five-year summary report. *CA Cancer J Clin.* 1982;32(4):194–225.
 24. Omranipour R, Kazemian A, Alipour S, Najafi M, Alidoosti M, Navid M, Alikhassi A, Ahmadinejad N, Bagheri K, Izadi S. Comparison of the accuracy of thermography and mammography in the detection of breast cancer. *Breast Care.* 2016;11(4):260–4.
 25. Keyserlingk JR, Ahlgren PD, Yu E, Belliveau N, Yassa M. Functional infrared imaging of the breast. *IEEE Eng Med Biol Mag.* 2000;19(3):30–41.
 26. Kennedy DA, Lee T, Seely D. A comparative review of thermography as a breast cancer screening technique. *Integr Cancer Ther.* 2009;8(1):9–16.
 27. FDA warns thermography should not be used in place of mammography to detect, diagnose, or screen for breast cancer: FDA Safety Communication. <https://www.fda.gov/medical-devices/safety-communications/fda-warns-thermography-should-not-be-used-place-mammography-detect-diagnose-or-screen-breast-cancer>. Date issued: 25 Feb 2019.
 28. Haleem A, Javaid M, Khan IH. Current status and applications of artificial intelligence (AI) in medical field: an overview. *Curr Med Res Pract.* 2019;9(6):231–7.
 29. Ratner M. FDA backs clinician-free AI imaging diagnostic tools. *Nat Biotechnol.* 2018;36:673–4.
 30. Petrone J. FDA approves stroke-detecting AI software. *Nat Biotechnol.* 2018;36(4):290.
 31. QuantX. Evaluation of automatic class III designation for QuantX. Available at: https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN170022.pdf
 32. Transpara. Radiological computer assisted detection/diagnosis software for lesions suspicious for cancer. Available at: <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K192287>
 33. Borchartt TB, Conci A, Lima RCF, Resmini R, Angel S. Breast thermography from an image processing viewpoint: a survey. *Signal Process.* 2013;93(10):2785–803.
 34. Singh D, Singh AK. Role of image thermography in early breast cancer detection-Past, present and future. *Comput Methods Prog Biomed.* 2020;183:105074.
 35. Prabha S, Suganthi SS, Sujatha CM. An approach to analyze the breast tissues in infrared images using nonlinear adaptive level sets and Riesz transform features. *Technol Health Care.* 2015;23(4):429–42.
 36. Pramanik S, Bhattacharjee D, Nasipuri M. Wavelet based thermogram analysis for breast cancer detection. In: 2015 International Symposium on Advanced Computing and Communication (ISACC). IEEE; 2015. p. 205–12.
 37. De Oliveira JPS, Conci A, Perez MG, Andaluz VH. Segmentation of infrared images: a new technology for early detection of breast diseases. In: 2015 IEEE International Conference on Industrial Technology (ICIT). IEEE; 2015. p. 1765–71.
 38. Kakileti ST, Manjunath G, Madhu HJ. Cascaded CNN for view independent breast segmentation in thermal images. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE; 2019. p. 6294–7.
 39. Raghavendra U, Rajendra Acharya U, Ng EYK, Tan J-H, Gudigar A. An integrated index for breast cancer identification using histogram of oriented gradient and kernel locality preserving projection features extracted from thermograms. *Quant InfraRed Thermogr J.* 2016;13(2):195–209.
 40. EtehadTavakol M, Sadri S, Ng EYK. Application of K-and fuzzy c-means for color segmentation of thermal infrared breast images. *J Med Syst.* 2010;34(1):35–42.
 41. Zadeh HG, Haddadnia J, Seryasat OR, Isfahani SMM. Segmenting breast cancerous regions in thermal images using fuzzy active contours. *EXCLI J.* 2016;15:532.
 42. Kakileti ST, Dalmia A, Manjunath G. Exploring deep learning networks for tumour segmentation in infrared images. *Quant InfraRed Thermogr J.* 2019;17:1–16.

43. Tavakol E, Mahnaz CL, Sadri S, Ng EYK. Analysis of breast thermography using fractal dimension to establish possible difference between malignant and benign patterns. *J Healthc Eng.* 2010;1:27–43.
44. Kakileti ST, Venkataramani K. Automated blood vessel extraction in two-dimensional breast thermography. In: 2016 IEEE International Conference on Image Processing (ICIP). IEEE; 2016. p. 380–4.
45. Kakileti ST, Madhu HJ, Manjunath G, Wee L, Dekker A, Sampangi S. Personalized risk prediction for breast cancer pre-screening using artificial intelligence and thermal radiomics. *Artif Intell Med.* 2020;105:101854.
46. Saniei E, Setayeshi S, Akbari ME, Navid M. A vascular network matching in dynamic thermography for breast cancer detection. *Quant InfraRed Thermogr J.* 2015;12 (1):24–36.
47. Silva LF, Sequeiros GO, Santos MLO, Fontes CAP, Muchaluat-Saade DC, Conci A. Thermal signal analysis for breast cancer risk verification. In: MedInfo. São Paulo: IOS press; 2015. p. 746–50.
48. Venkataramani K, Mestha LK, Ramachandra L, Prasad SS, Kumar V, Raja PJ. Semi-automated breast cancer tumor detection with thermographic video imaging. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE; 2015. p. 2022–5.
49. Roslidar R, Saddam K, Arnia F, Syukri M, Munadi K. A study of fine-tuning CNN models based on thermal imaging for breast cancer classification. In: 2019 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom). IEEE; 2019. p. 77–81.
50. Zuluaga-Gomez J, Al Masry Z, Benaggoune K, Meraghni S, Zerhouni N. A CNN-based methodology for breast cancer diagnosis using thermal images. *arXiv preprint arXiv.* 2019;1910.13757
51. Tello-Mijares S, Woo F, Flores F. Breast cancer identification via thermography image segmentation with a gradient vector flow and a convolutional neural network. *J Healthc Eng.* 2019;2019:1–13.
52. Flores JL, Gonzalez FJ, Cruz A, Navarro NE, Oceguera A. Automatic analysis of breast thermograms by convolutional neural networks. In: Applications of Digital Image Processing XLIII, vol. 11510. International Society for Optics and Photonics; 2020. p. 115101R.
53. Berz R, Schulte-Uebbing C. MammoVision (Infrared Breast Thermography) compared to x-ray mammography and ultrasonography. In: Diakides M, Bronzino JD, Peterson DR, editors. Medical infrared imaging: principles and practices. CRC Press; 2012. p. 12.1–12.12. <https://doi.org/10.1201/b12938-13>.
54. Kakileti ST, Venkataramani K, Madhu HJ. Automatic determination of hormone receptor status in breast cancer using thermography. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer; 2016. p. 636–43.
55. PROENG. Image processing and image analyses applied to mastology. 2012. <http://visual.ic.uff.br/en/proeng>
56. Sudhakar S, Manjunath G, Kakileti ST, Madhu H. Thermalytix: an advanced artificial intelligence based solution for non-contact breast screening. *Int J Med Health Sci.* 2018;12(2):48–51.
57. Wishart GC, Campisi M, Boswell M, Chapman D, Shackleton V, Iddles S, Hallett A, Britton PD. The accuracy of digital infrared imaging for breast cancer detection in women undergoing breast biopsy. *Eur J Surg Oncol (EJSO).* 2010;36(6):535–40.
58. Arora N, Martins D, Ruggerio D, Tousimis E, Swistel AJ, Osborne MP, Simmons RM. Effectiveness of a noninvasive digital infrared thermal imaging system in the detection of breast cancer. *Am J Surg.* 2008;196(4):523–6.
59. Hellgren RJ, Sundbom AE, Czene K, Izhaky D, Hall P, Dickman PW. Does three-dimensional functional infrared imaging improve breast cancer detection based on digital mammography in women with dense breasts? *Eur Radiol.* 2019;29(11):6227–35.
60. Kakileti ST, Madhu H, Manjunath G, Krishnan L, Sudhakar S, Rampakash HV. An observational study to evaluate the clinical efficacy of thermalytix for detecting breast cancer in symptomatic and asymptomatic women. *JCO Glob Oncol.* 2020;6:1472–1480.
61. Bratko I. Machine learning: between accuracy and interpretability. In: Learning, networks and statistics. Vienna: Springer; 1997. p. 163–77.
62. Vellido A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Comput & Applic.* 2020;32:18069–18083.
63. Kakileti ST. Machine Learning for Breast Cancer Diagnosis in Developing Countries. Maastricht: ProefschriftMaken; 2020. p. 216. <https://doi.org/10.26481/dis.20201109st>



AIM and Cervical Cancer

94

AI-Driven Early Detection of Cervical Cancer with Papsmear Analysis

Lipi B. Mahanta, Elima Hussain, and Kangkana Bora

Contents

Introduction	1318
Cervical Cancer and the Role of Pap Smear Test in Early Detection	1318
Pros and Cons of the Automated System of Diagnosis Concerning the Manual Diagnosis	1320
Challenges and Advances in Automation of Pap Smear Images	1321
Line of Action or AI-Driven Approaches for the Automation Task Using Pap-Smear Images	
Pap Smear Image Database Generation	1321
Ground Truth Labeling	1321
Automated Cell or Nuclei Segmentation from Whole Slide Image	1321
Feature Extraction and Feature Selection Methods	1324
Automated Binary or Multi-Class Classification of Pap Smear Images	1324
Summary	1325
References	1325

Abstract

Computer-assisted diagnosis has yielded disease prognosis and other clinical results for early diagnosis in a rapid manner. The applications of artificial intelligence for contributing immensely toward the automated diagnosis of cervical cancer including its staging have also come through many breakthrough development phases. Evidence in computer-assisted cervical cancer diagnosis has shown encouraging outcomes and warrant categorization of the pap smear images into its relevant classes based on The Bethesda System of Classification. Both advanced machine and deep learning techniques

Supplementary Information: The online version of this chapter (https://doi.org/10.1007/978-3-030-64573-1_253) contains supplementary material, which is available to authorized users.

L. B. Mahanta (✉) · E. Hussain
Institute of Advanced Study in Science and Technology,
Guwahati, India
e-mail: lbmahanta@iassst.gov.in;
lipimahanta@yahoo.co.in

K. Bora
Cotton University, Guwahati, India

have been explored to yield promising results in this subject. Instance and semantic segmentation, in-depth feature analysis based on shape, color, texture features, deep ensemble model classification, etc., are some of the prominent application areas using Pap smear images. The state-of-the-art techniques reveal 99.6% highest accuracy for Pap smear image classification. In a recent study, advanced simultaneous instance segmentation and classification have been achieved on Pap smear image using a proposed deep learning model.

Keywords

Cervical cancer · Pap smear test, Conventional method, Liquid-based cytology method · Instance segmentation · Deep learning classification task

Introduction

Cervical Cancer and the Role of Pap Smear Test in Early Detection

Cervical cancer is one of the many types of cancer which when detected early can be successfully cured at its initial stage. It is known to be the second most prevailing cancer among women because of its high incidence rate. Researchers have found that the human papillomavirus (HPV) plays the primary role of a cancer-causing agent in most of the cases. Early diagnosis through

periodic screening tests or HPV vaccination can thus reduce the increased risk of cervical cancer to a great extent. Very often women having early-stage cervical cancer do not undergo any symptoms and only the symptoms such as vaginal bleeding after sexual intercourse or menopause or after periods, pelvic pain after intercourse, etc., are experienced in advanced stages. This is the most common reason why cervical cancer is not detected in early stage in most developing or underdeveloped countries. Lack of awareness of screening tests is also a major reason. It is also a challenge to bring about an awareness of cervical cancer and early diagnosis among women belonging to developing countries like India, where personal fear and social taboo still prevails in the economically weaker section. Some of the associated risks factors include multiple sex partners, early sexual activity, multiple miscarriage or abortion, sexually transmitted infections (STI), taking birth control pills, etc. It is very important to have a routine Pap test and receive HPV vaccination to prevent the risk of cervical cancer. Prognosis treatment generally includes chemotherapy, surgery, or radiation therapy. Cervical cancer is of two types, namely, adenocarcinoma and squamous cell carcinoma based on types of cells in the endocervix and ectocervix. Endocervix involves glandular cells while ectocervix includes squamous cells. A schematic view is shown in Fig. 1.

The Papanicolaou or Pap smear test is one of the cervical screening techniques that facilitate the identification of precancerous cells in the cervix, which can be treated before they progress into

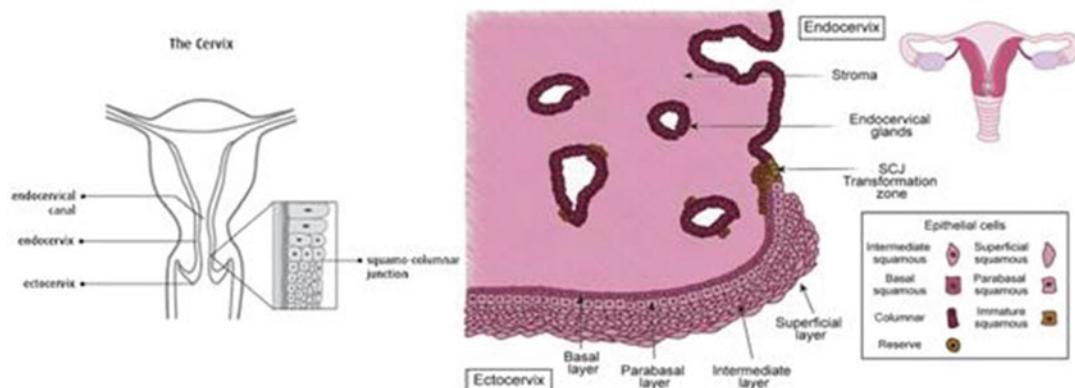


Fig. 1 Anatomy of human cervix highlighting ectocervix and endocervix regions

cancerous cells. Pap test is normally recommended for all sexually active women having age not less than 21 years. The specimen for smear is collected with the help of a speculum targeting the cells from the transformation zone, which is the highly sensitive region for abnormal cells to develop. The transformation zone features the location in the cervix where squamous epithelium cells meet columnar epithelium cells. The collected specimen is then transferred to a slide where the cells are spread over a certain small region. This region is then observed under a microscope by a pathologist, and any suspected cases are referred for a biopsy test for confirmation.

Conventional and liquid-based cytology (LBC) are the two available methods of the Pap smear test. The sample collection and technique for the preparation of the slides differ slightly for both the methods. Conventional Pap test involves taking the sample from the vagina with the help of a brush or spatula and then preparing and staining the smear on a glass slide manually by a technician who is trained for the same, with hematoxylin and eosin. The cell nuclei appear blue due to Hematoxylin stain whereas the extracellular particles and cytoplasm appear as pink as a result of eosin stain. The rest of the cellular structures take different shades, hues, or combinations of these colors. In the LBC method, the slides are prepared by a kit. It involves more steps such as the collected sample will be initially kept in a container having some additive fluids. This fluid helps in evacuating different types of debris such as mucus, blood cells, etc. The sample then

undergoes sedimentation and centrifugation at 2500 rpm for 5 min to break mucotic or blood molecules. This is again followed by staining using hematoxylin and Eosin (H&E) after which a uniform slide is obtained. An example related to sample preparation using LBC kit is shown in Fig. 2.

The two methods have their pros and cons. Whereas the LBC method is more efficient than conventional smears in providing more uniform and cleaner slides [1, 2], it involves a more elaborate and costly setup. In any underdeveloped or developing country, the economic viability of providing a certain type of medical test plays a pivotal role in its actual implementation, especially when it is utmost necessary to reach out to the prospective candidates as a screening method. However, apart from precancerous lesion detection, the LBC specimen can be used further for Human Papilloma Virus (HPV) testing, which is found to have the cancer precursor linking. Due to the manual method of preparation, conventional slides include many “debris” such as inflammatory cells, red blood cells, etc.

When adopting AI-driven automated methods for diagnosis, the choice of screening methods do not produce any diagnosis error but may require additional preprocessing and segmentation algorithms for removing the debris in case of conventional smears. Figure 3 illustrates the two cervical screening techniques schematically.

Based on the Pap test report, Squamous Intraepithelial Lesion (SIL) is reported following The Bethesda System (TBS) of Classification related



Fig. 2 (a) sample collection from patient (b) pap smear slide preparation using LBC kit (c) LBC kit

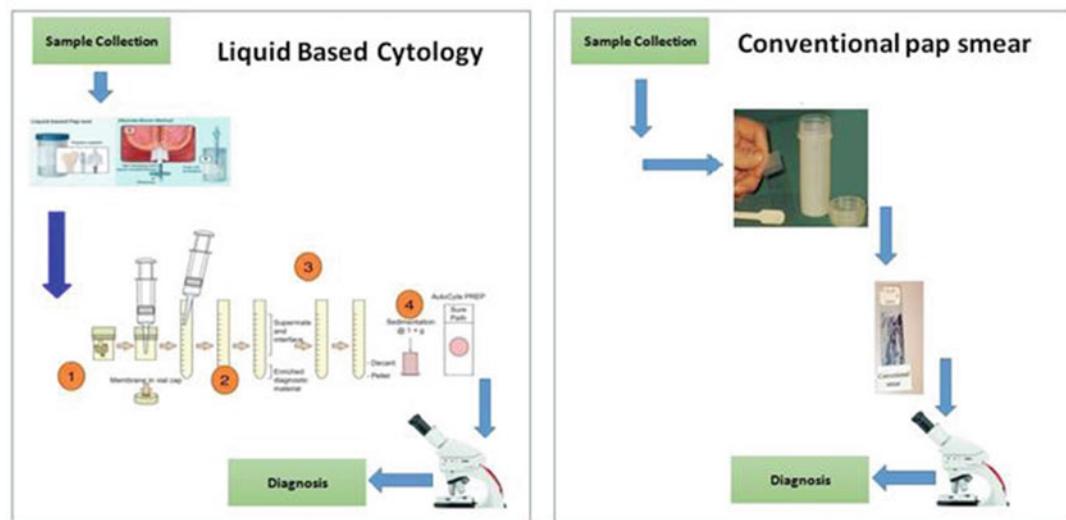


Fig. 3 Overview of Liquid-based cytology and Conventional Pap test

to cervix cancer [Nayar and Wilbur]. The system classifies SIL into the following:

1. Negative for intraepithelial malignancy (NILM)
2. Low-grade squamous intraepithelial lesion (LSIL) and
3. High-grade squamous intraepithelial lesion (HSIL)
4. Squamous cell carcinoma (SCC)

Apart from nuclear enlargement among the cervical cells, there are numerous cytological descriptions related to LSIL, HSIL, and squamous cell carcinoma (SCC) classes which are discussed well in the reference book by Gray et al. [3]. It has been established that 57% LSIL cases regress to normal while 32% LSIL cases progress to HSIL or carcinoma-in-situ and only 12% to invasive carcinoma cases.

Pros and Cons of the Automated System of Diagnosis Concerning the Manual Diagnosis

Class identification and quantification of cervical cancer by an experienced pathologist needs

rigorous observation of the individual cells, which is very tedious, time-consuming, and error-prone [4]. Considering the highly skewed ratio of pathologists/patients in most underdeveloped or developing nations, usually mass-screening following manual system involves huge processing and analysis time where the majority of cases come out to be under normal class. As an answer to this challenge, artificial intelligence (AI) via machine and deep learning have ushered into a much-needed automation aspect. Pap smear is one of the image-based screening tests, which can be integrated into automated image analysis or disease diagnosis software. Such an automated system may assist pathologists for rapid diagnosis. There are two commercially available cervical screening systems, namely, the Focal Point GS Imaging System by BD [5] and Thin Prep Imaging system by HOLOGIC, Inc. [6]. But such systems require skilled manpower and hence are not cost-effective and feasible during the mass-screening program. An automated system in such case which is more portable, robust, and low-cost can significantly contribute toward prognosis treatment. Pap smear test has helped to reduce the incidence rate of cervical cancer, so it is believed that the automation strategy will lead to a much better and rapid diagnosis.

Challenges and Advances in Automation of Pap Smear Images

Necrotic debris removal, poor staining quality or poor contrast images, and overlapped cytoplasm in Pap smear images are some of the prime research areas. Conventional pap smears often contain unwanted debris or artifacts such as inflammatory cells, red blood cells, etc., which require to be removed before automated classification. This is so because the shape, color, and texture features of individual cervical cells need to be studied and hence unwanted debris must be removed. There are some computer algorithms that address the solutions for debris removal. One such technique based on Maximally Stable Extremal Region (MSER) algorithm has been described by Bora et al. [7] for Pap smear images.

The unavailability of an indigenous dataset on pap smear images is also a major challenge for a researcher [8]. AI-driven methods delivered on indigenous dataset prove the consistency of real-time data. Another research challenge is to address overlapped cytoplasm or clustered nuclei problem for Pap smear images. Plissiti [9] and Song [10] have discussed techniques for tackling overlapped cytoplasm issue based on morphological reconstruction and graph cuts algorithms. Figure 4 illustrates general pipeline steps for automated diagnosis of cervical cancer.

Line of Action or AI-Driven Approaches for the Automation Task Using Pap-Smear Images

Pap Smear Image Database Generation

An image database plays the main role in image recognition. A well-curated image database serves as the strong pillar for any AI-based automation task. There is already a lack of indigenous medical image database in cervical cancer detection task and hence it is of concern area to prepare one manually and make it accessible to the research community. Among the very few available databases of conventional pap smear images include

the Sipakmed database by Plissiti et al. [11] and Pap smear benchmark database by Jantzen et al. [12]. Among LBC database include the liquid-based cytology pap smear dataset by Hussain et al. [13] and Cervix93 by Phoulady et al. [14]. All such database can be useful for AI or computer-assisted diagnosis, which requires interpretation of the images for different algorithms under computer vision and artificial intelligence. Figure 5 highlights pap smear sample images belonging to different classes.

In the case of Pap smear specimen, the whole slide images are first acquired using a digital microscope normally under $200\times$ or $400\times$ magnification. A minimum overlap of image sections is considered while capturing the images from the Pap smear slides. Ethical clearance has prime significance for conducting such studies.

Ground Truth Labeling

Ground truth labeling by an expert pathologist holds a key aspect for Region-of-Interest (ROI) selection. The ROI is manually confirmed by the expert, which is mandatory for accurate object recognition or image classification task. LabelMe and Icy are few existing software best suited for the ground truth labeling task.

The Region-of-Interest in a Pap smear image is mainly the cervical cells, namely, superficial, intermediate, basal, parabasal belonging to squamous epithelium layer. Apart from nuclear enlargement, numerous in-depth features related to these cells are analyzed for the final automation task.

Automated Cell or Nuclei Segmentation from Whole Slide Image

A segmentation algorithm plays a key role in extracting the Region-of-Interest. Any computer-assisted diagnosis system is significantly supported by accurate cell or nuclei segmentation. An efficient segmentation algorithm needs to consider the ROI in a whole slide pap smear image

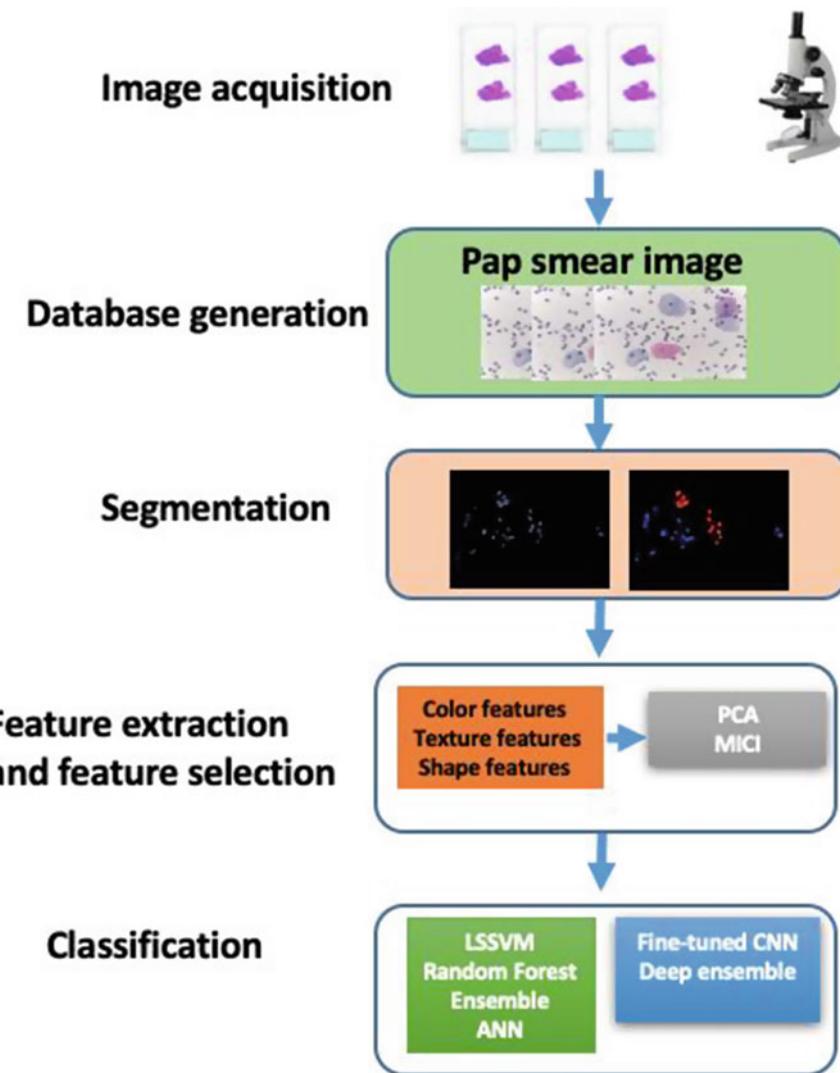


Fig. 4 General Steps in AI-driven applications for automated diagnosis

thus considering the removal of overlapped cytoplasm or background artifacts, etc., issues too.

The literature on automated segmentation using pap-smear images highlights few algorithms such as Adaptive Gradient Vector Flow [15], circular shape guided graph-based segmentation [16], Radiating Gradient Vector Flow Snake model [17], fuzzy c-means clustering [18], Particle Swarm Optimization algorithm [19], and modified Maximally Stable Extremal Region [7]. Further, there are algorithms for tackling overlapping cytoplasm issue along with segmentation

task on pap smear images such as integration of nuclei intensity or shape and texture feature by Song et al. [10] and Plissiti et al. [9]. However, all these algorithms follow different thresholding criteria often in pipelined structures. There are also some deep learning-based segmentation algorithms that follow pixel-wise classification. Deformable Multi-Path Ensemble Model based on a convolutional neural network by Zhao et al. [20], Mask_RCNN followed by conditional random field by Liu et al. [21], morphological convolutional neural network by Lin et al. [22],

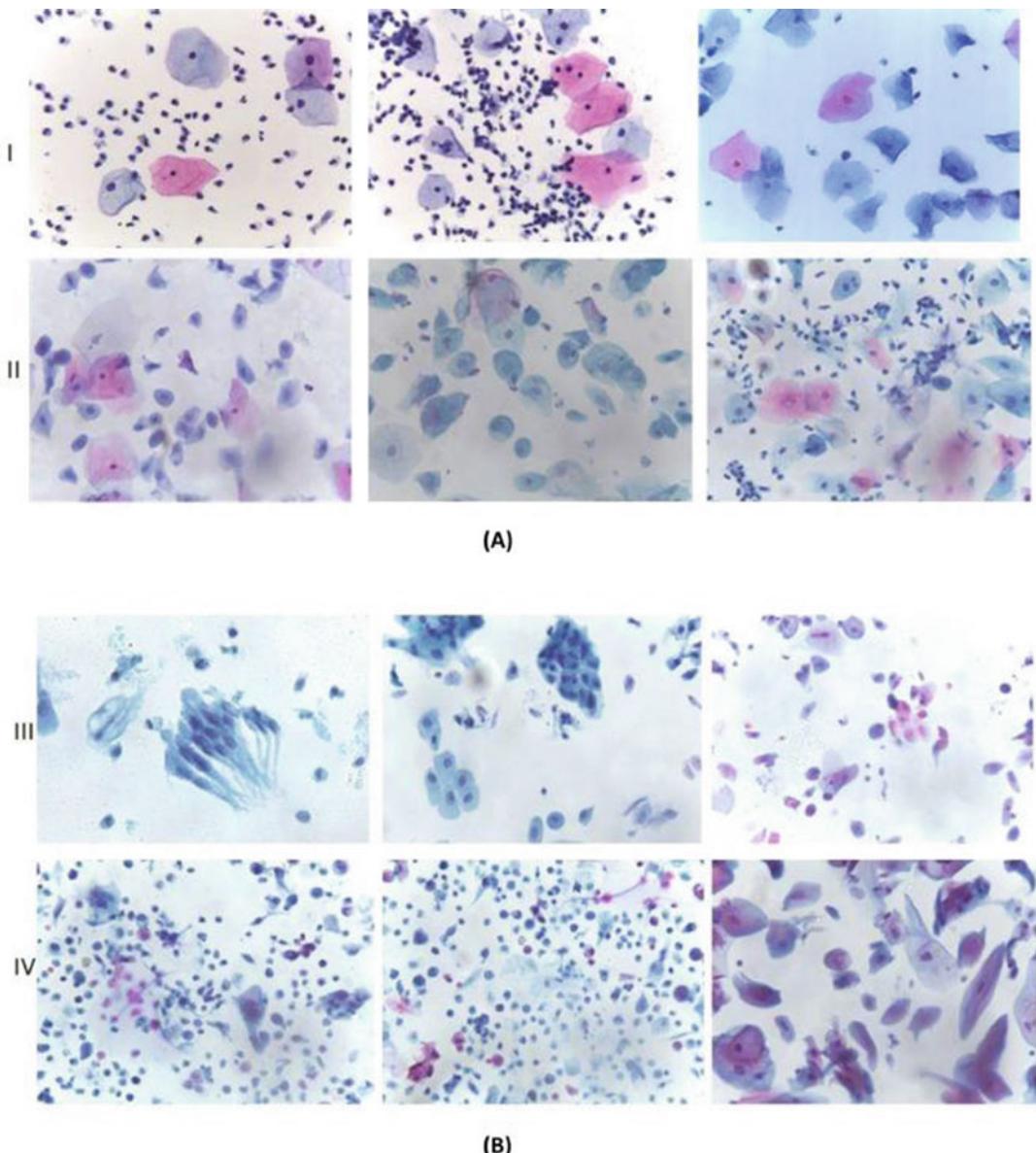


Fig. 5 Examples of whole slide Pap smear images belonging to (I) NILM, (II) LSIL, (III) HSIL, and (IV) SCC classes

Hussain et al. [23], and Song et al. [10] have addressed overlapping cell issue on pap smear images based on deep learning segmentation algorithms. Recent techniques such as simultaneous instance segmentation and classification have been explored by Hussain et al. on pap smear images. While conventional segmentation algorithms do not require input mask labeling, deep learning-based segmentation is solely dependent

on the input image and its associated mask. They require hard annotations for preparing the masks and may not be reliable for bulk computation tasks. In other words, the performance of segmentation algorithms is usually evaluated by Jaccard index or Zijdenbos Similarity Index to name a few, which compares the ratio of similarity between predicted segmented output and manual segmented output.

Feature Extraction and Feature Selection Methods

Feature extraction serves a main role for quantifying the dysplastic changes as well as differentiating the classes in whole slide Pap smear image. For this, in-depth analysis of both low-level and high-level clinical features related to the images are followed. Feature selection on the other hand involves selecting relevant feature for an increase in the classification accuracy. Generally, feature extraction and feature selection are core techniques under machine learning. While traditional machine learning classification models supports hand-engineered features, deep learning models can bring end-to-end classification outcome by learning both low-level and high-level features without selective feature engineering. The machine learning models rely on the extracted features to be fed as input parameters for the final classification task.

For quantification of pap smear images based on clinical features, one such work has been proposed by Bora et al. [7] based on cervical nuclei shape, color, and texture features. They have extracted the shape features using the MSER algorithm for quantifying the area, perimeter, eccentricity, and circularity of the cervical nucleus. The remaining texture and color features were extracted using Gray Level Co-occurrence Matrix (GLCM) and Ripple transform analysis. The literature on feature extraction process for Pap smear images highlights categorization based on different features used for quantifying the dysplastic changes. Work by Chen et al. [24] explains extraction of 13 features representing nuclear size, nucleo-cytoplasm (N/C) ratio, shape, and texture. Genctav et al. [8] have also performed in-depth feature analysis based on shape features. However, these two works failed to include color features. The color features using DFT and MPEG descriptors have been studied on pap smear images by Guan et al. [25] and Camargo et al. [26], respectively. But DFT can obtain frequency information and not time and has a bad convergence rate. MPEG descriptors neglect correlation from the pixels of neighboring blocks thus

producing the resultant image with undesired blocking artifacts. Ripple Type 1 transform proves to be a superior color feature extraction method than DFT and MPEG.

Feature selection enables reduced computational time and complexity by selecting a smaller subset of features thus retrieving the optimal characteristics of data. It also helps in the development of an efficient and compact classifier model. Bora et al. [27] have attempted one such unsupervised feature selection techniques on Pap smear images. The technique is based on measuring the similarity index using Maximal Information Compression Index among the feature vectors. The method proved to be better than Principal Component Analysis (PCA) as it is fast, do not require any search, and the feature subsets are not transformed. Initially, a feature set is partitioned into smaller clusters, which are having high similarity features but low inter-cluster similarity. Selection of a single feature from each cluster is done thus forming the resultant reduced subset.

Automated Binary or Multi-Class Classification of Pap Smear Images

Most of the machine learning models using the Pap smear images are reliant on pipelined segmentation or preprocessing algorithms, whereas deep learning model gives better classification output without segmentation. Literature works based on deep learning highlights binary and multi-class classification using single-celled (cropped-out single-cell images) as well as whole slide smear images. An accuracy of 99.6% has been achieved by Nirmaljith et al. [28] using a proposed deep convolutional neural network (CNN) for binary classification. There are many deep learning approaches for single-celled pap images such as patch-based CNN model by Gautam et al. [29], which reported 90% accuracy on binary classification and ConvNet architecture by Zhang et al. [30], which reported 98.3% accuracy. Based on experimental findings from such works, it can be observed that single-celled classification does not tackle real

problems arising from overlapped cytoplasms or clustered nuclei and hence cannot be applied to the whole-slide image. Work by Hussain et al. [31] reports an accuracy of 98.9% based on whole slide image using proposed deep ensemble models for multi-class classification. Among machine learning classifier models, Bora et al. [7] reported 96.5% accuracy using ensemble model (weighted majority voting scheme), Marinakis et al. [32] reported 96.8% accuracy using genetic algorithm-nearest neighbor, and Changkong et al. [18] reported 99.3% accuracy using artificial neural network. However, deep learning models require large-scale database for training the networks, which are not abundant in case of medical images. In such cases, data augmentation plays a crucial role in increasing the size of the image database. Inadequate data may cause overfitting or class-imbalance problem. Data augmentation supports predefined functions such as rotation, shearing, scaling, translation, cropping, zooming, flipping, etc.

Before starting the classification phase, it is very important to split the image database into training, validation, and testing sets. Strategies such as k-fold cross-validation, train-split strategy in the ratio of 80:20 can be followed here. For machine learning-based classifiers, the input feature sets extracted from feature analysis will be used whereas deep learning models prefer directly the image as input. Concepts such as sliding window protocol or image patch generation can be applied for reading the image initially. Classification using deep learning either involves transfer learning or building a CNN model from scratch. Such techniques do not require hand-engineered feature extraction. But the latter supports abundant large-scale database for training the model. Transfer learning using fine-tuned models can give promising results by pretraining a network on a large-scale natural database and then transferring the knowledge on our target domain. One such work on pap smear image analysis has been reported by Hussain et al. [31] where six fine-tuned deep learning models, namely, Alexnet, Resnet (resnet-50 and resnet-101), VGG-net (vgg-16 and vgg-19), and Googlenet have been explored.

Summary

Pap smear screening allows routine examination of the cervical lesion. A patient having low-grade or high-grade squamous epithelial can rapidly progress into squamous cell carcinoma if not detected early via routine screening tests. With the highly skewed ratio of pathologists to patients, it is very difficult to give rapid diagnosis result during mass screening campaigns. Artificial Intelligence can play a significant role in curbing out early diagnosis scenario and can ease the overall screening protocol.

References

- Cheung ANY, Szeto EF, Leung BSY, Khoo US, Ng AWY. Liquid-based cytology and conventional cervical smears: a comparison study in an Asian screening population. *Cancer*. 2003;99(6):331–5.
- Massad LS, Collins YC, Meyer PM. Biopsy correlates of abnormal cervical cytology classified using the Bethesda system. *Gynecol Oncol*. 2001;82(3):516–22.
- Gray W, Kocjan G. Diagnostic cytopathology. Churchill Livingstone, London, England; 2010.
- Betta PG, Andriola A, Donna A, Mollo F, Scelsi M, Zai G, et al. Malignant mesothelioma of the pleura. The reproducibility of the immunohistological diagnosis. *Pathol Res Pract*. 1997;193(11–12):759–65.
- Wilbur DC, Black-Schaffer WS, Luff RD, Abraham KP, Kemper C, Molina JT, et al. The Becton Dickinson focalpoint GS imaging system: clinical trials demonstrate significantly improved sensitivity for the detection of important cervical lesions. *Am J Clin Pathol*. 2009;132(5):767–75.
- Biscotti CV, Dawson AE, Dziura B, Galup L, Darragh T, Rahemtulla A, et al. Assisted primary screening using the automated ThinPrep Imaging System. *Am J Clin Pathol*. 2005;123(2):281–7.
- Bora K, Chowdhury M, Mahanta LB, Kundu MK, Das AK. Automated classification of Pap smear images to detect cervical dysplasia. *Comput Methods Prog Biomed* [Internet]. 2017;138:31–47. <https://doi.org/10.1016/j.cmpb.2016.10.001>.
- Gençtaş A, Aksoy S, Önder S. Unsupervised segmentation and classification of cervical cell images. *Pattern Recogn*. 2012;45(12):4151–68.
- Plissiti ME, Nikou C, Charchanti A. Automated detection of cell nuclei in Pap smear images using morphological reconstruction and clustering. *IEEE Trans Inf Technol Biomed*. 2011;15(2):233–41.
- Song Y, Zhang L, Chen S, Ni D, Li B, Zhou Y, et al. A deep learning based framework for accurate segmentation of cervical cytoplasm and nuclei. In: 2014 36th

- annual international conference of the IEEE Engineering in Medicine and Biology society (EMBC 2014); 2014. p. 2903–6.
11. Plissiti ME, Dimitrakopoulos P, Sfikas G, Nikou C, Krikoni O, Charchanti A. SIPAKMED: A new dataset for feature and image based classification of normal and pathological cervical cells in Pap smear images. In: IEEE International Conference on Image Processing (ICIP) 2018, Athens, Greece, 7–10 October 2018 [Internet]. 2018. Available from: <http://www.cs.uoi.gr/~marina/sipakmed.html>
 12. Jantzen J, Dounias G. The Pap smear benchmark. In: Proceeding of NYSIS-2006 Symposium [Internet]. 2006. Available from: <http://mde-lab.aegean.gr/index.php/downloads>
 13. Hussain E, Mahanta LB, Borah H, Das CR. Liquid based-cytology Pap smear dataset for automated multi-class diagnosis of pre-cancerous and cervical cancer lesions. Data Br [Internet]. 2020;30:105589. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S2352340920304832>
 14. Phouladhy HA, Moutan PR. A new cervical cytology dataset for nucleus detection and image classification (Cervix93) and methods for cervical nucleus detection. In: CVPR [Internet]. 2018. Available from: https://github.com/parham-ap/cytology_dataset
 15. Dong N, Zhao L, Wu A. Cervical cell recognition based on AGVF-Snake algorithm. Int J Comput Assist Radiol Surg [Internet]. 2019;14(11):2031–41. <https://doi.org/10.1007/s11548-019-01961-x>.
 16. Saha R, Bajger M, Lee G. Prior guided segmentation and nuclei feature based abnormality detection in cervical cells. In: Proceedings of 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE) 2019; 2019. p. 742–6.
 17. Li K, Lu Z, Liu W, Yin J. Cytoplasm and nucleus segmentation in cervical smear images using Radiating GVF Snake. Pattern Recog [Internet]. 2012;45(4): 1255–64. <https://doi.org/10.1016/j.patcog.2011.09.018>.
 18. Chankong T, Theera-Umpon N, Auephanwiriyakul S. Automatic cervical cell segmentation and classification in Pap smears. Comput Methods Prog Biomed [Internet]. 2014;113(2):539–56. <https://doi.org/10.1016/jcmpb.2013.12.012>.
 19. Marinakis Y, Marinaki M, Dounias G. Particle swarm optimization for pap-smear diagnosis. Expert Syst Appl. 2008;35(4):1645–56.
 20. Zhao J, Li Q, Li X, Li H, Zhang SL. Automated segmentation of cervical nuclei in Pap smear images using deformable multi-path ensemble model. In: ISBI. 2019. p. 1514–8.
 21. Liu Y, Zhang P, Song Q, Li A, Zhang P, Gui Z. Automatic segmentation of cervical nuclei based on deep learning and a conditional random field. IEEE Access. 2018;6:53709–21.
 22. Lin H, Hu Y, Chen S, Yao J, Zhang L. Fine-grained classification of cervical cells using morphological and appearance based convolutional neural networks. IEEE Access. 2019;7:71541–9.
 23. Hussain E, Mahanta LB, Das CR, Choudhury M, Chowdhury M. A shape context fully convolutional neural network for segmentation and classification of cervical nuclei in Pap smear images. Artif Intell Med [Internet]. 2020;107(May):101897. <https://doi.org/10.1016/j.artmed.2020.101897>.
 24. Chen Y-F, Huang P-C, Lin K-C, Lin H-H, Wang L-E, Cheng C-C, et al. Semi-automatic segmentation and classification of Pap smear cells. IEEE J Biomed Health Inform [Internet]. 2014;18(1):94–108. Available from: <https://ieeexplore.ieee.org/document/6482573/>
 25. Guan T, Zhou D, Xu C, Liu Y. A novel RGB Fourier transform-based color space for optical microscopic image processing. Robot Biomimetics [Internet]. 2014;1(1):16. Available from: <https://jrobo.springeropen.com/articles/10.1186/s40638-014-0016-1>
 26. Camargo LH, Diaz G, Romero E. Pap smear cell image classification using global MPEG-7 descriptors. Diagn Pathol [Internet]. 2013;8(Suppl 1):S38. Available from: <http://diagnosticpathology.biomedcentral.com/articles/10.1186/1746-1596-8-S1-S38>
 27. Bora K, Chowdhury M, Mahanta LB, Kundu MK, Das AK. Pap smear image classification using convolutional neural network. In: The ACM International Conference Proceeding Series (ICPS); 2016.
 28. Jith OUN, Harinarayanan KK, Gautam S, Bhavsar A, Sao AK. DeepCerv: Deep Neural Network for Segmentation Free Robust Cervical Cell Classification. In: First International Workshop, COMPAY2018, and 5th International Workshop, OMIA 2018, Held in Conjunction with MICCAI2018, Granada, 16–20 September 2018, Proceedings. 2018.
 29. Gautam S, Bhavsar A, Sao AK, Harinarayanan KK. CNN based segmentation of nuclei in PAP-smear images with selective pre-processing. In: Medical imaging: digital pathology. 2018. p. 32.
 30. Zhang L, Lu L, Nogues I, Summers RM, Liu S, Yao J. DeepPap: deep convolutional networks for cervical cell classification. IEEE J Biomed Health Inform. 2017;21(6):1633–43.
 31. Hussain E, Mahanta LB, Ray C, Kanta R. Tissue and Cell: A comprehensive study on the multi-class cervical cancer diagnostic prediction on Pap smear images using a fusion-based decision from ensemble deep convolutional neural network. Tissue Cell [Internet]. 2020;65:101347. <https://doi.org/10.1016/j.tice.2020.101347>.
 32. Marinakis Y, Dounias G, Jantzen J. Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification. Comput Biol Med. 2009;39(1): 69–78.



Artificial Intelligence in Infectious Diseases

95

Timothy Miles Rawson, Nathan Peiffer-Smadja, and Alison Holmes

Contents

Introduction	1328
Applications of Artificial Intelligence in Infectious Diseases	1329
Artificial Intelligence for the Identification of Microorganisms	1331
Microorganisms Detection, Identification, and Quantification	1331
Evaluation of Antimicrobial Susceptibility	1332
Diagnosis, Disease Classification, and Clinical Outcomes	1333
Artificial Intelligence for the Clinical Diagnosis and Management of Infectious Diseases	1333
Early Detection and Management of Sepsis	1333
Diagnosis of Infection	1334
Prediction Tools	1334
Antimicrobial Selection	1335
Artificial Intelligence in Surveillance and Infection Prevention	1336

T. M. Rawson · A. Holmes (✉)
Centre for Antimicrobial Optimisation, Imperial College
London, London, UK

Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, Imperial College London, Hammersmith Hospital, London, UK
e-mail: timothy.rawson07@imperial.ac.uk;
alison.holmes@imperial.ac.uk

N. Peiffer-Smadja
Centre for Antimicrobial Optimisation, Imperial College London, London, UK

Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, Imperial College London, Hammersmith Hospital, London, UK

Université de Paris, Inserm, IAME, Paris, France
e-mail: nathan.peiffer-smadja@inserm.fr;
n.peiffer-smadja@imperial.ac.uk

Challenges and Limitations in the Development and Application of Artificial Intelligence in Infectious Disease Management	1336
Development, Implementation, and Adoption	1337
Conclusion and Recommendations for Artificial Intelligence in Infectious Diseases	1338
References	1338

Abstract

The management of infectious diseases lends itself to the application of artificial intelligence. The treatment of infection is complex, requiring the consideration of a large number of dynamic variables to inform decision-making. This includes considering organism, host, and drug factors in the context of local disease epidemiology and potential long-term consequences of anti-infective use, such as the development of antimicrobial resistance. The heterogeneity of clinical presentation caused by the same pathogen means that in many cases there is a paucity of data available to guide decision-making in real time, with individualized decisions made based on the individual patient and available data. Within this chapter we explore current applications of artificial intelligence in (i) the laboratory detection of microorganisms, (ii) the clinical diagnosis and management of infectious diseases, and (iii) the surveillance of infection. This chapter will not address other potential areas for the application of AI in infectious diseases that include anti-infective drug development, targeting infection prevention activity, and public health decision-making. We highlight potential future directions for AI in infectious diseases within the areas explored by this chapter and current barriers to wider adoption of such systems.

Keywords

Infectious disease · Microbiology · Antimicrobial resistance · Supervised machine learning · Unsupervised machine learning · Clinical decision support systems

Introduction

The management of infectious diseases is practiced ubiquitously across medicine in all countries and health economies. Infection can be the primary reason that someone is admitted to hospital or can complicate other diseases, for example, postoperative wound infection, *Clostridioides difficile* (*C. difficile*) infection following antimicrobial exposure, or neutropenic sepsis following chemotherapy.

Infectious disease management lends itself to the application of artificial intelligence (AI). Management of any infectious disease requires a large number of dynamic variables to be considered. During the course of an infection these variables are constantly changing and interacting in different ways. The clinician must consider individual patient factors, such as their age, physiological status, the site of the infection, their comorbidities, previous infections, and other medical treatments that they are receiving. They must consider the causative organisms characteristics, such as its susceptibility to antibiotics, its pathogenicity, and its ability to evade the immune system. They must consider the characteristics of any treatment that they recommend, such as the ability of an antimicrobial to penetrate the site of infection, the concentration that certain doses will achieve following the pharmacokinetics and pharmacodynamics (PK/PD) of the drug, and potential side effects to the patient. Additionally, infectious disease specialists must consider the impact of the infection and its treatment on the wider population, through the potential transmission to other individuals and the propagation of antimicrobial resistance (AMR).

Often, limited research and clinical data are available to support the evidence-based diagnosis and management of infection. Consequently, the use of antimicrobials for the treatment of infection is often inappropriate, which includes delays in appropriate antimicrobial use and overuse. Inappropriate use of antimicrobials is a key modifiable driver of AMR.

Adoption of electronic clinical decision support systems (CDSS) and the integration of hospital datasets have enhanced clinician decision-making by providing person-specific and population-level data at the point decisions are made. When used, CDSS improve the quality and safety of healthcare provided [1]. Global uptake of electronic health record (EHR), computer prescribing order entry (CPOE) systems, and mobile technology are generating greater volumes of data to support clinical decision-making in infection management. The development of new diagnostics, powerful processing capabilities, and AI provide opportunities to utilize available data in a more precise manner. This may facilitate better decision-making in infection management through the delivery of individualized, evidence-based recommendations.

Within the field of infectious diseases, a number of AI-based systems have been developed, evaluated, and reported. These range from AI tools to support laboratory diagnosis of pathogens to empirical antimicrobial selection, and to the better use of data for surveillance of healthcare-associated infections [2, 3].

Figure 1 outlines some of the potential applications of AI technology to the field of infectious diseases. Training AI to perform data cleaning functions and to deal with missing data presents exciting opportunities to improve the use of routine EHR data for decision-making. Advances in computer processing capabilities and AI algorithms could facilitate real-time prediction. AI may enhance our ability to make individualized treatment decisions for the management of infection, even when there is a limited evidence base to support decision-making. AI may also facilitate surveillance, including for the unintended

consequences of decisions, such as side effects and the potential propagation of AMR, making healthcare safer. However, many barriers still also remain in the development, implementation, and adoption of AI technologies in infectious disease management.

Within this chapter, we explore the rationale for AI use in infectious diseases. We report the current state-of-the-art AI supporting (i) the laboratory identification of microorganisms, (ii) the clinical diagnosis and management of infection, and (iii) the surveillance of diseases. We also highlight future potential applications of different types of AI and provide examples of successes and challenges to date.

Applications of Artificial Intelligence in Infectious Diseases

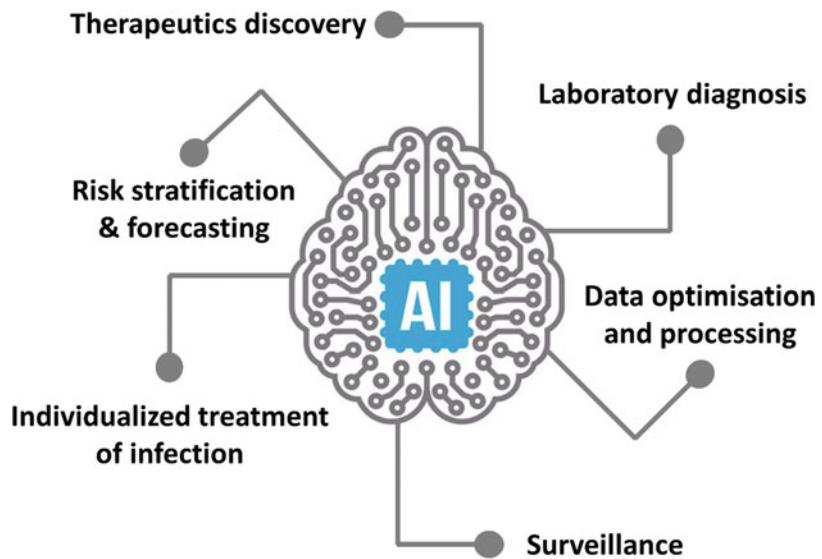
Infectious disease management is a complex process, requiring consideration of a large number of variables (Fig. 2).

Optimal treatment of the individual with infection must consider host, organism, and drug factors. This must be balanced against potential future impact on other patients and future populations through the transmission of infection and propagation of AMR.

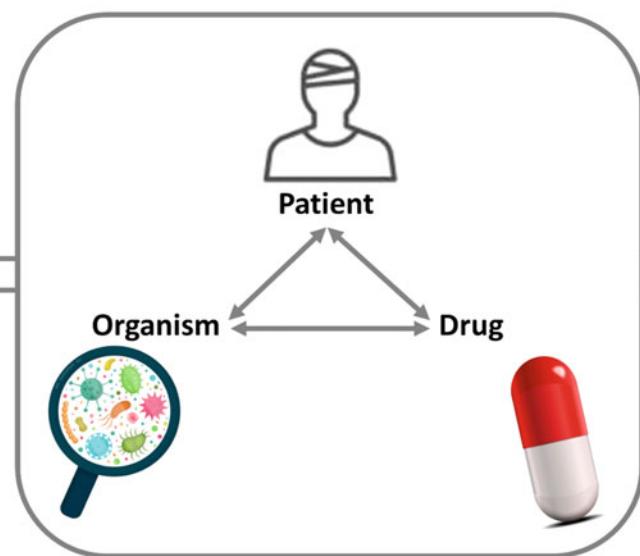
Current data available to support clinical decision-making must be carefully interpreted. For example, the microbiology laboratory can only provide in vitro data on drug susceptibility that must be contextualized to the infection that the clinician is treating. Whole-genome sequencing may provide a genotypic picture of an organism, but the phenotype being expressed in vivo may be different. Drug PK-PD in the context of the individual patient must also be considered, with different target site concentration and risk of resistance development taken into account.

Despite infectious diseases being managed across all specialities of healthcare, training in infectious diseases is limited at the undergraduate and postgraduate level [4, 5]. For example, primary care physicians in the UK who are

Fig. 1 Potential applications of artificial intelligence technology in the field of infectious disease management. Artificial intelligence can be applied to a large number of fields within infectious diseases including the laboratory diagnosis of infection, clinical management of disease, and surveillance. Furthermore, AI can support the discovery of new antimicrobial agents, better sorting of routine healthcare data and forecasting at individual and population levels



Current evidence



Transmission dynamics



Population surveillance data



Future, global impact

Fig. 2 Interactions that must be considered when making decisions for infectious disease management. The management of infection requires consideration of a large number of variables that are constantly changing and interacting in different ways. This includes consideration of the individual patient, pathogen, and antimicrobial therapy that is

being prescribed. Additionally, the clinician must take into account factors such as local epidemiology of disease and drug resistance, the evidence available to support decision-making, risk of transmission of infection to others, as well as the potential long-term consequences of prescribing therapy on antimicrobial resistance

responsible for approximately 75% of all antibiotic prescribing are only expected to complete 2/1368 (0.14%) curriculum learning points on

antibiotic prescribing and its consequences (e.g., AMR) during their postgraduate training [4]. The application of AI to low-middle income settings

may also support decision-making in settings where infections are managed by nonphysician healthcare professionals, such as nurse-led clinics, which rely on syndromic algorithms to guide decisions on the diagnosis and management of infectious diseases.

Given the limited amount of formal training and the complexity of decision-making, the ability to provide a greater level of decision support in infectious disease management is crucial. The application of AI offers potential solutions to this broad challenge. Within this chapter, we will explore how AI has been applied to support decision-making in the diagnosis, management, and surveillance of infectious diseases. We will describe some of the current barriers to implementation and adoption and outline recommendations for the future development and evaluation of AI platforms in infectious diseases.

Artificial Intelligence for the Identification of Microorganisms

The identification and characterization of microorganisms is a cornerstone of infectious disease management. Until recently, the laboratory process had remained unchanged for nearly 100 years relying on traditional microbiology techniques to identify and characterize microorganisms causing infection. Within the last decade, advances such as matrix-assisted laser desorption ionization–time-of-flight (MALDI-TOF) mass spectrometry (MS), whole-genome sequencing (WGS), and the access to microbiota data by next-generation sequencing has driven a shift in processes within the laboratory [6]. This has generated greater amounts of clinically relevant data to support clinical decision-making during the diagnosis and management of infectious diseases. The application of AI in the microbiology laboratory provides the opportunity to better harness data from other technologies, to reduce the time between sampling and microbiological diagnosis. AI can also improve the use of traditional methods, such as microscopic imaging and bacterial culture [3, 7].

The application of AI in the laboratory can support the diagnosis of infection, the identification, and quantification of micro-organisms, and the analysis of antimicrobial susceptibility. A recent review identified 97 AI applications designed to support microorganism detection, identification, and quantification, evaluate antimicrobial susceptibility, help target diagnosis, and support disease classification and prediction of clinical outcomes [3]. Most of these AI technologies focused on diagnosis of bacterial infection (85%). Diagnosis of parasitic, viral, and fungal infection has also been reported.

Current AI technology in the microbiology laboratory often aims to facilitate the automation of repetitive tasks with the clinical microbiologists required to validate their results. Future applications for AI in this setting aim to provide systems capable of analyzing complex and high-dimensional data, such as that generated by next-generation sequencing. They also may support the development of lab-free, point-of-care technologies.

Microorganisms Detection, Identification, and Quantification

Microbiology laboratories are moving towards (semi-) automated systems. AI can be used to optimize the identification of bacteria by direct analysis of clinical samples, using volatile organic compounds for diabetic foot infection or tuberculosis, 16S ribosomal DNA PCR on respiratory samples for lower respiratory tract infections (LRTI), or fluorescent microscopic images in tuberculosis [3]. Recently, the use of Raman optical spectroscopy of a single bacterial colony in suspension coupled with a convolutional neural network (CNN) was used to identify a single species among 30 bacterial and yeast species within hours after clinical sampling [8]. This technology could support much more rapid turnaround times for organism detection and reporting than current methods employed in routine practice.

For parasitic disease, image recognition software coupled with algorithms such as artificial neural networks (ANN) and k-nearest neighbor

classification (KNN) have been applied to microscopic images of blood smears to detect *Plasmodium* species and microscopic images of fecal samples to identify intestinal helminths [9, 10]. This approach can be used to support diagnosis in settings where skilled laboratory staff are not available and also support screening of large volumes of samples in settings where staff are present.

In the virology laboratory, the application of ANN and random forest analysis has been used to analyze metagenomics sequences with classifiers prespecified to identify new putative viral sequences and support taxonomic classification from large datasets [11].

Evaluation of Antimicrobial Susceptibility

Determination of antimicrobial susceptibility is a core function of the microbiology laboratory. In vitro antimicrobial susceptibility is used to guide appropriate treatment choices for infectious disease management. It can also be used to identify patients with potentially transmissible healthcare-associated infections (HCAI), such as MRSA and carbapenemase-producing Enterobacteriaceae (CPE). While a range of methods are used to screen for and determine significant antimicrobial susceptibility patterns, the mainstay of most laboratories remains the disc diffusion method [12]. Newer methods of determining organism resistance, such as whole-genome sequencing and real-time PCR, are available. However, these only provide an organism genotype. As organisms, such as CPE, do not always express their mechanism of resistance, the reliance of genotype alone can be misleading.

Inferring the phenotypic antibiotic susceptibility pattern of microorganisms from genomic data is one of the main areas that can be supported by the application of AI [13]. AI can be applied when knowledge of the precise mutational event associated with antibiotic resistance is not complete and therefore cannot be used in genotype-to-phenotype studies [13]. Support vector machine classification (SVM) has been used to analyze *Pseudomonas*

aeruginosa gene expression (RNAseq) and the presence or absence of resistance genes demonstrating a high accuracy for predicting resistance to the broad-spectrum antibiotics meropenem and tobramycin [14]. However, worse performance for predicting resistance to other antibiotics, such as ceftazidime (81% for predicting resistance and 83% for predicting susceptibility), was observed [14]. Similar approaches have been described for *C. difficile*, *Escherichia coli*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, and *Elizabethkingia* species [3]. One study looked at combined genotypic and phenotypic data to rapidly identify the susceptibility profile of bacterial pathogens from RNAseq data directly from positive blood cultures, developing a random forest model classifier to support analysis [15]. Another study conducted a supervised classification and regression tree analysis with clinical outcome data from 58 patients to determine the optimal pyrazinamide MIC for *Mycobacterium tuberculosis* susceptibility breakpoint [16]. Similar application of AI technology has been used in the field of HIV to explore mutations associated with antiretroviral agent resistance. This could help to find unexpected mutations associated to the evolution of viral sequences in HIV and thereby pave the way for new diagnostic markers that could replace standard genotypic analyses [3].

Improving the speed and cost of current laboratory techniques has also been addressed through the development of AI technology. Using Raman spectroscopy and CNN, it was possible to differentiate methicillin-sensitive and methicillin-resistant *Staphylococcus aureus* strains (MSSA and MRSA) with high accuracy a few hours after sampling [8]. The development of amplification and melting curve analysis (AMCA) using intercalating dyes as part of a digital PCR is one novel approach to this problem [17]. Supervised machine learning can be applied to the analysis of real-time amplification data and thermodynamic melting profiles to accurately classify genotypic resistance patterns, such as colistin resistance in multidrug resistance gram-negative bacteria [17]. Further application of AMCA, supported by supervised machine learning approaches, is driving the development of point-of-care

diagnostics that do not require traditional laboratory capacity and facilities available.

Diagnosis, Disease Classification, and Clinical Outcomes

To support the diagnosis of infectious diseases, one group combined microbiological and host transcriptome data in tracheal aspirates using a rules-based model (RBM) to differentiate patients with proven LRTI ($n = 26$) and patients with noninfectious acute respiratory failure ($n = 18$) with a 100% sensitivity and 87% specificity [18]. ANN have been developed to analyze the odor (volatile organic compounds) of various clinical samples in order to diagnose urinary tract infections, acute chronic obstructive pulmonary disease exacerbations, or active tuberculosis [3].

AI technology linked with microbiology laboratory data has been used to predict the severity of disease such as respiratory syncytial virus infection. Indeed, analysis with AI identified a 84-gene expression pattern that could discriminate between children with mild respiratory syncytial virus (RSV) infection from those with severe RSV disease [19]. The identified signature was independently validated in other cohorts and could serve to develop prognostic tests for the management of RSV disease. Other similar AI technologies have been evaluated to predict the recurrence of *C. difficile* [20] or the severity of tuberculosis from WGS data or gut microbiota data.

Extreme gradient boosting has been used to analyze urine microscopy, culture, and sensitivity reports over one year in a single microbiology laboratory and develop a method of screening urine samples and classifying those that can be discarded because of a low probability of significant microbiological growth. Using 212,554 urine sample reports, including clinical and epidemiological data, the authors stated that such an application could reduce the workload by 41% [21]. However, the implementation of the system and the evaluation of its actual impact are still in progress.

Artificial Intelligence for the Clinical Diagnosis and Management of Infectious Diseases

The clinical diagnosis and management of infectious disease is challenging. It requires the combination of laboratory microbiology, clinical information, and a wider understanding of patient and regional trends in disease and epidemiology. Given that most infections are managed by healthcare professionals who are not experts in infection management, there is a need to provide expert clinical support for decision-making. Currently, support for decision-making takes the form of guidelines, clinical prediction rules (determined using traditional statistics), and clinician experience. This leads to wide variations in practice, many of which are often not evidence based.

AI offers dynamic, individualized methods of supporting decision-making for infection management based on the individual patient. Current AI technologies have been developed to focus on a range of clinical outcomes, such as the prediction of sepsis in critical care, the diagnosis of TB or surgical site infection, the prediction of virological success of antiretroviral therapy, and the selection of an antibiotic regimen.

One review identified 60 unique AI tools designed to support decision-making for clinical diagnosis and management of infection [2]. AI was used to support the diagnosis of infection (20/60; 33%), the early detection or stratification of sepsis (18/60; 30%), the prediction of response to antimicrobial therapy (13/60; 22%), the presence of antibiotic resistance (4/60; 7%), and the choice of antibiotic regimen (3/60; 5%).

Early Detection and Management of Sepsis

Sepsis is defined as a dysregulation of the body's immune response to an infective trigger that often drives multiple organ dysfunction. Steps critical to reduce mortality require early detection, determination of the probable source of the infection (e.g., urinary, central nervous system, and chest), and commencement of appropriate treatment. In

particular, delay or inappropriate prescription of antibiotics in sepsis is associated with higher rates of death and long-term morbidity [22]. The application of AI for the diagnosis and management of sepsis may be able to improve outcomes by providing tailored, evidence-based decision support to clinicians.

For development of AI to support decision-making in sepsis, critical care datasets such as the Medical Information Mart for Intensive Care (MIMIC) have provided rich data from which algorithms can be developed and tested [23]. Many AI tools designed to support the diagnosis and management of sepsis include a large number of variables, often richly measured and stored electronically within critical care databases. These include individual patient vital signs, laboratory data, demographic information, medical history, therapeutic data, and radiology plus specialist investigations. In some cases, unstructured clinical data is also added to these models.

Markov decision processes have been applied to intensive care data as part of the “AI Clinician” to explore whether AI can make individualized treatment decisions in the context of sepsis [21]. In a large validation dataset, sepsis treatment decisions matching those recommended by the AI clinician were associated with superior outcomes to divergent treatment decisions [21]. This supports the hypothesis that AI can be used to compute large volumes of data, rapidly accumulating experience that is greater than individual humans can acquire, and apply these to real-world situations to support optimal decision-making.

Diagnosis of Infection

Beyond sepsis, supporting the appropriate diagnosis of infection remains a significant challenge for clinicians. While blood stream infection and sepsis will often present with recognizable signs and symptoms, infections such as urinary tract infection or postoperative complications can often be difficult to differentiate from non-infectious signs and symptoms. These tend to drive overprescribing of antibiotics, which has a significant impact on the selection of AMR.

AI to support the diagnosis or exclusion of infection has been widely explored in a range of conditions. These range from the diagnosis of tuberculosis (TB) in outpatient settings to the diagnosis of bacterial infection in hospitalized patients, or the distinction between bacterial and viral meningitis [2].

Supervised and unsupervised machine learning has been explored for the prediction of events. This has predominantly focused on predicting the likelihood of infection using clinically available data. For example, TREAT explored the ability of causal probabilistic networks to predict the likelihood of blood stream infection and causative organism [24, 25]. Other examples include the use of decision tree classifiers using binary classification applied to blood test parameters to predict the diagnosis of *Chlamydia pneumoniae* [26] and hepatitis B/C virus [27]. SVM have been used to predict the likelihood of a patient having a positive microbiological result within 48 h of routine blood tests, which successfully inferred the likelihood of infection in a prospective evaluation of a cohort of patients admitted to a hospital in London, UK [28].

Prediction Tools

Wide variation in host, organism, and drug factors plus traditional delays in the determination of infection type and susceptibility profile mean that antimicrobial therapy is often empiric during the initial phase of therapy. Once targeted therapy can be confidently delivered, the large number of remaining unknown factors mean that confidence in outcome, determination of length of therapy, and final outcome/long-term sequelae cannot always be predicted. Wide heterogeneity and lack of experimental data means that the clinician will often rely on experience or unsupported, but widely accepted “clinical rules,” such as treatment for 5, 7, 10, or 14 days.

In the outpatient setting, ANN have been applied to predict treatment success including the virological response to HIV and HCV therapy [2]. These tended to analyze treatment history, viral genotype, and medical history. Treatment

history was limited to previous HIV or HCV therapies but not to other drugs that the patient might have been taking. SVM, random forest, and least absolute shrinkage and selection operator (LASSO) regression have been used to predict treatment outcome in TB using demographic data (e.g., education level and homelessness), TB history including constitutional symptoms, treatment, and structured data from the chest radiograph (e.g., size of cavity) [29].

In the hospital setting, *C. difficile* colitis and disease recurrence have been predicted [2]. AI including decision trees and regression techniques have been explored for personalized prediction of baseline antibiotic resistance in UTI [30] and treatment history of patients with positive blood cultures to predict baseline susceptibility to ampicillin, ceftriaxone, and gentamicin [31].

Antimicrobial Selection

Antimicrobial selection is traditionally guided by local prescribing policy. This is determined using local antibiograms, clinician experience, and probable causative organisms for the type of infection. AI can provide individualized recommendations, taking into account a much greater number of factors than the ones currently used to determine antimicrobial guidelines.

Current AI systems for antimicrobial selection have focused on choice of the optimal individual antibiotic regimen in bacterial infections and anti-retroviral treatment for HIV [2]. To date, these systems only focus on selection of the optimal antimicrobial agent, but have not explored route, dose, or duration of treatment. These systems have tended to use data routinely available in the hospital setting.

A notable example of AI used in clinical infectious diseases, with high-quality data to support its application, is the TREAT system. TREAT used causal probabilistic networks [24, 25]. A cluster randomized control trial was undertaken across three hospitals with the primary outcome measures of appropriateness of empirical prescribing. A secondary analysis of 180-day

survival following treatment was also explored. For influence of AI on the appropriateness of empirical therapy compared to detected organisms' sensitivity, TREAT demonstrated a 9% improvement in appropriateness of prescribing. However, once findings were adjusted for medical ward clustering and site, using multivariate regression, the findings did not reach significance. This may have been partly due to under powering of the study, or financial and time constraints. Furthermore, assessment of 180-day survival demonstrated a significant benefit from the use of AI on per-protocol analysis (6% increase in survival, $p = 0.04$). Statistical significance was not reached on intention-to-treat analysis. These findings suggest that clinical uptake of interventions may have been a contributing factor, along with appropriate powering of the cluster RCT.

Causal probabilistic networks, or Bayesian networks, are an attractive option for the utilization of machine learning in medicine. They facilitate the incorporation of qualitative and quantitative variables to model uncertain knowledge [32]. However, a major problem of these types of knowledge-based systems are that they require the construction of large, complex decision trees. In the case of TREAT, this comprised over 6000 nodes [24, 25]. The complexity of these type of system leads to problems transferring them into clinical practice as there is wide heterogeneity in infrastructure and data availability [32]. Therefore, at present these types of tool are challenging to develop, implement in practice, and require large amounts of information and technical skill to maintain [34].

Alternative methods have been explored to support antimicrobial selection in practice. Case-based reasoning (CBR) uses knowledge-based systems. CBR aims to solve a new problem by adapting a previously successful solution to the current problem encountered [34]. CBR aims to address many of the challenges associated with knowledge-based systems, including:

1. CBR does not require a defined model like causal probabilistic networks. Therefore, data collection simply relies on the extraction of case histories.

2. Implementation of CBR requires the identification of significant features within a case, as opposed to creating an explicit model.
3. CBR facilitates large volumes of information to be managed by applying database techniques. Furthermore, it provides greater flexibility when working with sparse or incomplete datasets.
4. CBR learns through cases that it acquires, which makes maintenance of such systems easier than model-based systems.

CBR has been used widely in the field of medical decision-making including antibiotic decision-making in intensive care [33, 35–37], radiology, psychiatry, chronic disease management, hepatology, and cancer.

CBR for infectious diseases has been applied and prospectively tested in clinical practice [38]. The use of a CBR CDSS recommended narrower-spectrum antimicrobials, while maintaining a similar appropriateness of prescriptions recommended by clinicians in practice [38]. Experimental evaluation of this type of approach is yet to be described.

Future applications of AI in clinical diagnosis and management of infectious disease must be developed in diverse health settings, including primary care and low-middle income (LMIC) settings that are currently underrepresented. Developed tools must be embedded and integrated with current electronic health record technology specific to the clinical setting that is it deployed within. Future studies should aim to report clinical outcomes following sustainable use in routine clinical care and focus on both patient outcomes (i.e., mortality) and long-term benefits on AMR and HCAI.

Artificial Intelligence in Surveillance and Infection Prevention

Infection prevention and control (IPC) is as important as the laboratory diagnosis and clinical management of infection in promoting optimal patient outcomes and preventing long-term consequences of AMR and HCAI. The use of AI in IPC has generally been applied to surveillance and diagnosis/detection of infections and outbreaks. The use of AI in the microbiology laboratory

have already been described above. With the application of point-of-care technology, this provides a mechanism for leapfrogging many of the current barriers to the application of surveillance in low-middle income settings [39].

Surveillance of infection, such as HCAI, lends itself to the application of AI. Surveillance of HCAI requires the analysis of data from multiple different databases to prospectively monitor trends, identify outbreaks, predict future events, and detect networks of transmission [40]. For common HCAI, such as *C. difficile*, AI has been applied to EHR data to predict the risk of infection and respond to the dynamic nature of healthcare [40].

To support screening for potentially transmissible diseases, deep learning image recognition has been applied to the analysis of chest radiograph to support the diagnosis/detection of pulmonary tuberculosis (TB). Automated systems may greatly support current mechanisms of IPC. Furthermore, their application to screening of images in settings where there is limited specialist radiologist input may facilitate the streamlining of cases to those only where there is diagnostic uncertainty/a high likelihood of TB being present.

Current challenges in the application of AI to surveillance systems involves the lack of standardized data being collected for a wide range of infections. This data is often not rich enough for significant systems to be applied outside of individual institutions or regions making translatability challenging. Moreover, variations in microbiology laboratory practice make standardized tools that work across different regions difficult to develop.

Challenges and Limitations in the Development and Application of Artificial Intelligence in Infectious Disease Management

While AI has the potential to enhance infectious disease management, a number of barriers remain to its adoption. These can be classified at the micro (individual health professional) and macro (health system and beyond) barriers. A common concern with AI is the widespread, unregulated training of these systems during their development, as well as

the potential of unsupervised learning upon implementation. Insufficient attention to the human-computer interface as part of clinical decision support at the design phase means systems are ergonomically weak.

At the macrolevel, there is currently very little guidance and regulation governing the training and development of AI in infectious diseases and medical microbiology. As with many other fields, dataset size, generalizability of training sets, and evaluation of systems before deployment are often left to the discretion of the developer, with safety and translatability to other clinical situations often unclear. This lack of regulation has led to concerns about adoption of such tools, providing a major barrier to successful implementation. At the microlevel the unsupervised nature with which systems can potentially continue to learn in the absence of human oversight means that errors input into the system may be propagated and go unnoticed unless supervised and curated, leading to patient societal harm. Currently, there is little thought to how these potential unintended consequences of AI will be monitored for and how systems will respond to them when they are identified.

Current media attention around the use of AI has attracted some strong opinions in terms of safety in healthcare. From the patient perspective, confidence in computer-driven decisions regarding healthcare is mixed. In a recent study exploring public confidence in computer-driven decisions, citizens reported that while computer-supported decision-making had potential benefits in improving accuracy, interaction and oversight from a medical professional remained important [41]. This was reported by participants to be based on the realization that human interaction cannot simply be based on evidence, but must also take the social and cultural context in which the decision is made into account. For antibiotic prescribing, many potential end users are currently uncomfortable about the way that decision support will be implemented [41], especially when the system may tend towards a more societal perspective than that of the individual, as is often the case when decisions are made for infection management.

Development, Implementation, and Adoption

Many of these described barriers can be addressed through the engagement of healthcare professionals, patients, and carers early in and during the development of such tools. AI development emphasizes participatory approaches because, in essence, it aims to simulate capacities of human intelligence. However, while codesign of the system to improve design and the functionality aspects are important, it may not be enough. Codesign must also alleviate concerns and help to promote greater transparency in how the recommendations being made by the system have been developed and are utilized to augment human decision-making. This will promote better understanding of the role of these tools in healthcare and help foster the realization that although human subjectivity can be removed from prediction, it cannot be removed from the practice of evidence-based medicine. This is because the individual, social, and microbiological context in which a decision is made requires the human decision-maker.

For the clinical decision-maker, AI tools remain notoriously difficult to interpret. While visualization tools may highlight the combination of variables that led an algorithm to make a particular prediction, healthcare professionals must be aware that, like humans, AI can easily be affected by systematic biases (e.g., scanning device, patient's age, etc.). Special pedagogical efforts must thus be made in both scientific reports and in the clinics to keep a healthy skepticism when it comes to AI results. Second, the lack of large healthcare, clinical, imaging, and genetic public repositories leads each institution to locally develop its own analytical pipeline on its own small dataset, which significantly limits the generalizability of the results. While this issue is not specific to AI technology, the ability of modern algorithms to encompass heterogeneous datasets should drive us to both share the de-anonymized raw data used in each clinical study and favor the development of large cohorts from which AI technology can be applied. Furthermore, AI technology must be evaluated using randomized control methodology using clear outcome

measures such as mortality to further support their application in practice.

Conclusion and Recommendations for Artificial Intelligence in Infectious Diseases

AI technology has great potential to support better decision-making in the management of infectious diseases. This ranges from enhancing the ability of the microbiology laboratory to better and more rapidly identify pathogens to helping clinicians make optimal, individualized treatment decisions for specific infections. To recognize the potential of AI in infectious diseases many of the general limitations in the development and implementation of AI, discussed elsewhere in this book, must be addressed. Additionally, for infectious diseases, the development of AI technology must take into account the application of tools in low-middle income settings and utilize datasets that allow the system to control for many of the systemic biases that are observed in current approaches. Implementation of AI technology must be supported by robust, randomized control trial evidence that looks at hard outcomes, such as the impact of AI on patient mortality, taking into account the behavioral and nontechnical factors that play a large role in the diagnosis and management of infectious diseases. Finally, with the ever-expanding generation of electronic data and methods of being able to utilize this to support decision-making, we must also consider the ethics of failing to use this data to support decision-making if a clear benefit from the application of AI technology is demonstrated.

Acknowledgments The authors would like to acknowledge Dr. Pantelis Georgiou, Dr. Pau Herrero, and Dr. Bernard Hernandez from the Center for Bioinspired technology, Imperial College London, UK. They also acknowledge the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in healthcare-associated infection and antimicrobial resistance at Imperial College London in partnership with Public Health England and the NIHR Imperial Patient Safety Translational Research Center. The Department of Health and Social Care funded Center for Antimicrobial Optimization (CAMO), Imperial College London, provides state-

of-the-art research facilities and consolidates multidisciplinary academic excellence, clinical expertise, Imperial's NIHR/Wellcome funded Clinical Research Facility (CRF), and partnerships with the NHS to support and deliver innovative research on antimicrobial optimization and precision prescribing. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the UK Department of Health.

References

1. McGregor JC, Weekes E, Forrest GN. Impact of a computerized clinical decision support system on reducing inappropriate antimicrobial use: a randomized controlled trial. *J Am Med Inform Assoc.* 2006;13:378–84.
2. Peiffer-Smadja N, Rawson TM, Ahmad R, Buchard A, Pantelis G, Lescure FX, et al. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. *Clin Microbiol Infect.* 2020;26:584–95.
3. Peiffer-Smadja N, Delliére S, Rodriguez C, Birgand G, Lescure FX, Fourati S, et al. Machine learning in the clinical microbiology laboratory: has the time come for routine practice? *Clin Microbiol Infect.* 2020;26:1300–9.
4. Rawson TM, Butters TP, Moore LSP, Castro-Sánchez E, Cooke FJ, Holmes AH. Exploring the coverage of antimicrobial stewardship across UK clinical postgraduate training curricula. *J Antimicrob Chemother.* 2016;71:3284.
5. Castro-Sánchez E, Drumright LN, Gharbi M, Farrell S, Holmes AH. Mapping antimicrobial stewardship in undergraduate medical, dental, pharmacy, nursing and veterinary education in the United Kingdom. *PLoS One.* 2016;11:1–10.
6. Buchan BW, Ledeboer NA. Emerging technologies for the clinical microbiology laboratory. *Clin Microbiol Rev.* 2014;27:783–822.
7. Weis CV, Jutzeler CR, Borgwardt K. Machine learning for microbial identification and antimicrobial susceptibility testing on MALDI-TOF mass spectra: a systematic review. *Clin Microbiol Infect.* 2020;26:1310–7.
8. Ho C-S, Jean N, Hogan CA, Blackmon L, Jeffrey SS, Holodniy M, et al. Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning. *Nat Commun.* 2019;10:4927.
9. Park HS, Rinehart MT, Walzer KA, Chi J-TA, Wax A. Automated detection of *P. falciparum* using machine learning algorithms with quantitative phase images of unstained cells. *PLoS One.* 2016;11: e0163045.
10. Yang YS, Park DK, Kim HC, Choi MH, Chai JY. Automatic identification of human helminth eggs on microscopic fecal specimens using digital image processing and an artificial neural network. *IEEE Trans Biomed Eng.* 2001;48:718–30.

11. Bzhalava Z, Tampuu A, Bała P, Vicente R, Dillner J. Machine learning for detection of viral sequences in human metagenomic datasets. *BMC Bioinformatics*. 2018;19:336.
12. EUCAST. Disk diffusion method for antimicrobial susceptibility testing-antimicrobial susceptibility testing EUCAST disk diffusion method. 2021.
13. Ruppé E, Cherkaoui A, Lazarevic V, Emonet S, Schrenzel J. Establishing genotype-to-phenotype relationships in bacteria causing hospital-acquired pneumonia: a prelude to the application of clinicalmetagenomics. *Antibiotics*. 2017;6:1–15.
14. Khaledi A, Weimann A, Schniederjans M, Asgari E, Kuo T-H, Oliver A, et al. Predicting antimicrobial resistance in *Pseudomonas aeruginosa* with machine learning-enabled molecular diagnostics. *EMBO Mol Med*. 2020;12:e10264. <https://doi.org/10.15252/emmm.201910264>.
15. Bhattacharyya RP, Bandyopadhyay N, Ma P, Son SS, Liu J, He LL, et al. Simultaneous detection of genotype and phenotype enables rapid and accurate antibiotic susceptibility determination. *Nat Med*. 2019;25: 1858–64.
16. Gumbo T, Chigutsa E, Pasipanodya J, Visser M, van Helden PD, Sirgel FA, et al. The pyrazinamide susceptibility breakpoint above which combination therapy fails. *J Antimicrob Chemother*. 2014;69:2420–5.
17. Moniri A, Miglietta L, Malpartida-Cardenas K, Pennisi I, Cacho-Soblechero M, Moser N, et al. Amplification curve analysis: data-driven multiplexing using real-time digital PCR. *Anal Chem*. 2020;92:13134–43.
18. Langelier C, Kalantar KL, Moazed F, Wilson MR, Crawford ED, Deiss T, et al. Integrating host response and unbiased microbe detection for lower respiratory tract infection diagnosis in critically ill adults. *Proc Natl Acad Sci*. 2018;115:E12353–62.
19. Jong VL, Ahout IML, van den Ham H-J, Jans J, Zaaraoui-Boutahar F, Zomer A, et al. Transcriptome assists prognosis of disease severity in respiratory syncytial virus infected infants. *Sci Rep*. 2016;6:36603.
20. Staley C, Kaiser T, Vaughn BP, Graizer CT, Hamilton MJ, Rehman T u, et al. Predicting recurrence of *Clostridium difficile* infection following encapsulated fecal microbiota transplantation. *Microbiome*. 2018;6:166.
21. Burton RJ, Albur M, Eberl M, Cuff SM. Using artificial intelligence to reduce diagnostic workload without compromising detection of urinary tract infections. *BMC Med Inform Decis Mak*. 2019;19:171.
22. Paul M, Shani V, Muchtar E, Kariv G, Robenshtok E, Leibovici L. Systematic review and meta-analysis of the efficacy of appropriate empiric antibiotic therapy for sepsis. *Antimicrob Agents Chemother*. 2010;54: 4851–63.
23. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035.
24. Paul M, Andreassen S, Nielsen AD, Tacconelli E, Almanasreh N, Fraser A, et al. Prediction of bacteremia using TREAT, a computerized decision-support system. *Clin Infect Dis*. 2006;42:1274–82.
25. Paul M, Nielsen AD, Goldberg E, Andreassen S, Tacconelli E, Almanasreh N, et al. Prediction of specific pathogens in patients with sepsis: evaluation of TREAT, a computerized decision support system. *J Antimicrob Chemother*. 2007;59:1204–7.
26. Richardson A, Hawkins S, Shadabi F, Sharma D, Fulcher J. Enhanced laboratory diagnosis of human Chlamydia pneumoniae infection through pattern recognition derived from pathology database analysis. In: Third IAPR international conference on pattern recognition in bioinformatics (PRIB 2008). Monash University; 2008. p. 227–34.
27. Richardson AM, Lidbury BA. Infection status outcome, machine learning method and virus type interact to affect the optimised prediction of hepatitis virus immunoassay results from routine pathology laboratory assays in unbalanced data. *BMC Bioinformatics*. 2013;14:206.
28. Rawson TM, Hernandez B, Moore LSP, Blandy O, Herrero P, Gilchrist M, et al. Supervised machine learning for the prediction of infection on admission to hospital: a prospective observational cohort study. *J Antimicrob Chemother*. 2019;74:1108–15.
29. Sauer CM, Sasson D, Paik KE, McCague N, Celi LA, Sánchez Fernández I, et al. Feature selection and prediction of treatment failure in tuberculosis. *PLoS One*. 2018;13:e0207491.
30. Yelin I, Snitser O, Novich G, Katz R, Tal O, Parizade M, et al. Personal clinical history predicts antibiotic resistance of urinary tract infections. *Nat Med*. 2019;25:1143–52.
31. Oonsivilai M, Mo Y, Luangasanatip N, Lubell Y, Miliya T, Tan P, et al. Using machine learning to guide targeted and locally-tailored empiric antibiotic prescribing in a children's hospital in Cambodia. *Wellcome Open Res*. 2018;3:131.
32. Daly R, Shen Q, Aitken S. Learning Bayesian networks: approaches and issues. *Knowl Eng Rev*. 2011;26:99–157.
33. Pantic M. Introduction to machine learning & case-based reasoning. Imperial College London; 2006. <https://ibug.doc.ic.ac.uk/media/uploads/documents/courses/syllabus-CBR.pdf>.
34. Watson I, Marir F. Case-based reasoning: a review. *Knowl Eng Rev*. 2017;94:327–54.
35. Godo L, Puyol-Gruart J, Sabater J, Torra V, Barrufet P, Fàbregas X. A multi-agent system approach for monitoring the prescription of restricted use antibiotics. *Artif Intell Med*. 2003;27:259–82.
36. Schmidt R, Gierl L. Case-based reasoning for antibiotics therapy advice: an investigation of retrieval algorithms and prototypes. *Artif Intell Med*. 2001;23:171–86.
37. Heindl B, Schmidt R, Schmid G, Haller M, Pfaller P, Gierl L, et al. A case-based consiliarius for therapy recommendation (ICONS): computer-based advice for calculated antibiotic therapy in intensive care medicine. *Comput Methods Biomed*. 1997;52:117–27.
38. Rawson TM, Hernandez B, Moore LSP, Herrero P, Charani E, Ming D, et al. A real-world evaluation of a case-based reasoning algorithm to support

- antimicrobial prescribing decisions in acute care. *Clin Infect Dis.* 2020. <https://doi.org/10.1093/cid/ciaa383>.
39. Okeke IN, Feasey N, Parkhill J, Turner P, Limmathurotsakul D, Georgiou P, et al. Leapfrogging laboratories: the promise and pitfalls of high-tech solutions for antimicrobial resistance surveillance in low-income settings. *BMJ Glob Heal.* 2020;5: e003622.
40. Fitzpatrick F, Doherty A, Lacey G. Using artificial intelligence in infection prevention. *Curr Treat Options Infect Dis.* 2020;12:135–44.
41. Rawson TM, Ming D, Gowers SA, Freeman DM, Herrero P, Georgiou P, Cass AE, O'Hare D, Holmes AH. Public acceptability of computer-controlled antibiotic management: an exploration of automated dosing and opportunities for implementation. *J Infect.* 2019;78(1):75–86



Thomas Lefèvre and Cyrille Delpierre

Contents

Introduction	1342
Present Advances	1344
Collecting Data for AI Development and Developping AI for Data Collection	1344
Disease and Health Outcomes Surveillance	1345
Pharmacovigilance	1346
Sentiment Analysis and the Use of Social Media as Alternative Data Source and Data Processing Method in Epidemiology	1346
From Data to Decision: Data- and Model-Driven Knowledge and Decision in Epidemiology	1347
Potential Trends and Future Challenges	1348
The Challenges Related to Data	1348
The Challenges Related to Epidemiology as an Explanatory Science: Statistical Inference and Causality	1349
The Challenges Related to the Use Made of Epidemiology by Decision-Makers and the Involvement of Private Actors	1350
Conclusion	1350
References	1350

Abstract

T. Lefèvre (✉)
IRIS Institut de Recherche Interdisciplinaire sur les enjeux Sociaux, UMR8156 CNRS – U997 Inserm – EHESS – Université Sorbonne Paris Nord, Paris, France

Department of Forensic and Social Medicine, AP-HP, Jean Verdier Hospital, Bondy, France
e-mail: thomas.lefeuvre@univ-paris13.fr

C. Delpierre
CERPOP, Center for Epidemiology and Research in POPulation Health, Université de Toulouse, Inserm, UPS, Toulouse, France
e-mail: cyrille.delpierre@inserm.fr

John Last defined epidemiology as “The study of the distribution and determinants of health-related states or events in specified populations, and the application of this study to the control of health problems.” It underscores that epidemiologists are not concerned only with disease but with health-related events, and that ultimately epidemiology is committed to control of disease. Initially focused on the disease, the objects of investigation in epidemiology now correspond to any

factor that may influence the state of health of the human being, i.e., biological, clinical factors, in relation to the physical, mental, and social environment. Regardless of the field considered and the type of epidemiology (clinical or population based) referred to, the basic brick of epidemiology remains the data. The data must be as valid and precise as possible to ensure validity and reliability of results. The use of artificial intelligence and its methods can occur at different levels and in several areas of epidemiology. At present, we can consider three main use cases. First, AI can add to a long tradition of using more or less sophisticated observational data analysis methods. It has a role to play in causal inference. Second, AI can intervene at the stage of reconciling and structuring siled and varied data sources. Finally, AI can simply bring new ways of exploring and using data, such as sentiment analysis applied to social media. These three use cases are found, in practice, often intermingled and do not necessarily meet in isolation from each other. The private sector (intermediation platforms) and policy makers are the two other actors involved in the forms that AI uses in epidemiology will take.

Keywords

Epidemiology · Public health · Risk factor · Clinical epidemiology · Population health · Health care policy · Artificial intelligence · Machine learning · Structure learning · Causal inference

Introduction

The origin of epidemiology lies in the idea first expressed more than 2000 years ago by Hippocrates that environmental factors can influence the occurrence of diseases [1]. The word epidemiology comes from the Greek word “epi,” meaning on or over, “demos,” meaning people, and “logos,” meaning the study of. In other words, the word epidemiology has its roots in the study of what happens to a population. Multiple

definitions have since been proposed. John Last, in the *Dictionary of Epidemiology* [2], has defined epidemiology as “The study of the distribution and determinants of health-related states or events in specified populations, and the application of this study to the control of health problems.” Last’s definition underscores that epidemiologists are not concerned only with disease but with “health-related events,” and that ultimately epidemiology is committed to control of disease. As such, epidemiology constitutes a basic discipline of public health since it aims to quantify the health status of populations in order to identify its causes and to propose interventions to improve the health of the population concerned and to evaluate them. Depending on the population concerned, a distinction can be made between clinical epidemiology (“clinical research”), which is concerned with populations of “patients” with a view to improving the medical management of these patients, and population epidemiology (“classical” epidemiology), which is concerned with the general population with a view to developing public health strategies.

The different fields of epidemiology are characterized by the nature of the questions to be answered and the methods used to answer them. Classically, a distinction is made between: i) descriptive epidemiology, the purpose of which is to describe the frequency and distribution of health phenomena or health determinants in populations, according to human, spatial, and temporal characteristics. The purpose here is to estimate surveillance indicators such as prevalence, incidence, and survival; ii) analytical epidemiology, the purpose of which is to identify and estimate the link between exposure to certain factors and the subsequent occurrence of disease (or health event). The object here is the search for determinants, to study the patterns, the “causes”; and iii) evaluative epidemiology, which aims to test the effectiveness of a health program or policy intervention on the disease or health condition under study.

Initially focused on the disease, the objects of investigation in epidemiology now correspond to any factor that may influence the state of health of the human being (i.e., biological, clinical factors,

in relation to the physical, mental, and social environment).

Regardless of the field considered and the type of epidemiology (clinical or population based) referred to, the basic brick of epidemiology remains the data. The data must be as valid and precise as possible to ensure that they measure what they are supposed to measure (validity) and limit measurement error (reliability).

Whether it is to describe a phenomenon, to understand mechanisms, or to make inferences, evidence in epidemiology is based on the protocolized/standardized production of data. Protocols specify the volume and the scope of the data to be collected. The logic consists of working with a limited set of data that are precisely characterized and traced, and even certified. It is therefore the quality of the data, and what it allows for an increase in generality, that is prioritized. It is a question of specifying a set of criteria upstream of the investigation in order to characterize the population studied, the values investigated, and to ensure that the measures are likely to respect its characteristics.

In clinical epidemiology, these data most often come from clinical and biological data collected routinely in medical records, for example, or ad hoc through the implementation of specific studies (clinical trials, cohorts), or from specific systems that ensure the validity and exhaustiveness of the data, such as cancer registries. In population epidemiology, data can also come from trials, registries, or cohorts, which have the advantage of including all the desired data but for a restricted population (the one included in the study) and from data routinely collected in different systems, such as medico-administrative data (reimbursement of expenses/medical acts), administrative data (civil status, taxes, family allowances, retirement, etc.), and environmental data (air pollution), which are likely to provide information on the living conditions of a very large part of the population (or even the entire population). The limitation of these data is that they have not been constructed to study the health of populations, which raises questions about their validity and reliability, and these different databases were not constructed with the intention of being

interoperable with other databases. Access to these databases, particularly at the regulatory level, is also an issue.

Once these data have been collected, they are then modeled. Epidemiology has always used numerical methods to produce results, whether to help describe, understand, or evaluate parameters hidden from observation, and to help with experimentation [3]. Concerning the modeling used, we can distinguish classical regression methods such as multiple linear, logistic regression, and survival models, and also more sophisticated methods (causal approach), some of which are now labeled “machine learning” (hierarchical classification, factorial analysis, component analysis, canonical correlation analysis, and Bayesian networks). The definition, collection, and analysis of data is therefore at the heart of epidemiology, which makes it a discipline particularly concerned by the development of what is called massive data and artificial intelligence.

The development of data digitization, including in the field of health (biology, clinical, and imaging), connected tools, and social networks, opens a new era of availability of massive, varied data, never known before. The term big data is thus used, which is classically associated with the 3Vs, namely volume (which accounts for the massive use of data), variety (which relates to the heterogeneity of content), and velocity (which is associated with the speed of data processing) [4]. Big data may therefore be defined as the rapid processing of massive heterogeneous data that calls upon approaches, some of which are part of what is known as artificial intelligence (AI).

These massive data are opportunities to better describe, understand, and intervene on the health of populations, and therefore a tool for epidemiology and its development. As stated by Flahault et al., the main objective of the data science “is to develop mathematical models and computational solutions able to reason and interpret massive amount of data where typically the information sought is quite sparse” [5] and to make it possible to apprehend and make intelligible an ever-increasing amount of information that humans cannot condense and understand in a meaningful way [6].

But AI's immediate usefulness is not guaranteed, however: Electronic health records were predicted to transform primary care for the better, but led to unanticipated outcomes and encountered barriers to adoption [7, 8]. AI could also harm, for example, by exaggerating racial, class, or gender biases if models are built with massive but biased data that are not representative of the entire population, or used with new populations for whom performance may be poor [9].

Present Advances

The use of artificial intelligence and its methods can occur at different levels and in several areas of epidemiology. At present, we can consider three main use cases:

- Epidemiology has a long tradition of using more or less sophisticated methods of observational data analysis. The so-called AI methods extend and enrich this tradition, and can be compared to more classical methods, when they are used to pursue the same goals, for example, to identify associations between an exposure and a health outcome.
- The profusion of data, as well as the diversity of their collection format (voice, text, sound, etc.), and also the real-time nature of the collection of this data mean that AI methods can impose themselves as natural reference methods, for uses or analyses significantly different from what epidemiology has produced until now.
- Finally, epidemiology is often seen as intervening at the stage of the analysis of data and their interpretation. This same diversity of sources and types of data, their heterogeneity, and their distribution in many more or less partitioned systems require methods upstream of the analyses. These methods are typically AI methods, and make it possible to collect, extract, and structure data according to different models, so that they can then be used by epidemiologists.

These three use cases are found, in practice, often intermingled and do not necessarily meet in isolation from each other. We present the

illustration through several examples below: the use of AI for the organization of data, the surveillance of diseases and health outcomes, pharmacovigilance, sentiment analysis of social networks, and finally the use of data and model-driven approaches [10].

Collecting Data for AI Development and Developing AI for Data Collection

Data are the building blocks useful for epidemiology, but also for any learning of an AI or a model in general. Thus, it is necessary to be able, on the one hand, to collect and annotate data to train the AI algorithms which will be useful for epidemiology, and on the other hand, the AI algorithms will also be able to be used to help this data collection.

AI to Collect, Classify, and Structure Data

AI can be used to match data with each other from separate databases, especially when the data is anonymized and a probabilistic matching needs (data linkage) to be performed. This may be the case, for example, between cohort databases, electronic health records (EHR), medico-administrative databases, or even open data. A particular use of AI named federated learning is developing more and more, in order to overcome a double problem: that of the distributed nature of data and that of the robust learning of algorithms.

Thus, AI can intervene to reconcile heterogeneous sources or formats toward a single and shared structured model: This can be the case from electronic health records, for example [11].

The use of interoperability standards as an intermediate or complementary solution, overlaying already available data sources, is developing to allow a shared research framework. This is the case with the common data model from the OMOP initiative (Observational Medical Outcomes Partnership [11]). Finally, a third possibility, that of not bringing together the data in one and the same database, for reasons of properties, rights, and privacy: It becomes possible to carry out the training of AI algorithms useful for epidemiology by “distributing” these algorithms to the different sources (for example: in each

participating hospital). These techniques are called federated learning. They have already been used in several cases, such as the prediction of side effects related to drugs, apart from electronic health records [12]. Xu et al. define federated learning as “a mechanism of training a shared global model with a central server while keeping all the sensitive data in local institutions where the data belong, provides great promise to connect the fragmented healthcare data sources with privacy-preservation” [13]. This emerging technique seems to be able to overcome the majority of the obstacles encountered when learning AI from distributed sources, and the performances of the algorithms thus obtained seem to be able to be almost similar to those obtained from a centralized database: The Sheller et al.’s experience showed that federated learning based on the exploitation of data from ten establishments made it possible to obtain models of quality similar to those obtained by centralizing the data (99% of the model quality with centralized data) [14].

AI to Reconstruct or Virtualize Experimental Designs

The use of multivariable models in epidemiology, applied to observational data, tends to neutralize as much as possible the influence of confounding factors, and therefore to approximate a controlled experimental design, as in clinical trials. What we could not impose at the time of data collection, we try to recover by the sophistication of the analyses and models used. One promise of AI is to go even further, by recreating from massive amounts of data, databases that virtually reproduce the conditions of a controlled experimental design. We are talking about virtual randomized controlled trials [15, 16]. These approaches remain for the moment essentially theoretical.

Disease and Health Outcomes Surveillance

Public health surveillance may be defined as the ongoing systematic collection, analysis, and interpretation of data, closely integrated with the timely dissemination of the resulting information

to those responsible for preventing and controlling disease and injury [17]. Multiple sources of data have become usable for public health surveillance, for example, mobile phones, online searches, social media, credit card transactions, wearable and ambient sensors, electronic health records (EHRs), medico-administrative records, and pharmacy sales.

It can be expected that the “rapid development of data science, encompassing big data and AI, and the exponential growth of accessible and highly heterogeneous health-related data should enable health authorities to respond readily to health crises” [18]. By using traditional approaches (for example, from paper-based medical records), it could take weeks to find out that an infection was emerging somewhere in the world. The access and use of data now generated almost continuously from hospitals or other medical facilities and devices constitutes theoretically an opportunity to understand, predict, and combat diseases at unprecedented speed.

Some studies have already shown the interest of using a large number of queries on Internet search engines to detect earlier influenza epidemics in the USA [19]. The pandemic of COVID-19 is an ongoing example of the importance of data to face the epidemic. This pandemic shows us how much data on mobility via cell phones, use of public transport, leisure facilities, and social networks can be used to control the epidemic. The same is true for wastewater data, which can be used to locate places that are more exposed to the virus. All of these data enable the development of more accurate prediction models, which in turn can be used to better guide public decisions.

We can be confident that public health surveillance will use this type of approach in addition to the systems currently in place for the surveillance of infectious diseases but also chronic diseases. However, it is important to keep in mind that these tools will not replace existing surveillance systems. It has been shown that Google Flu Trends, which estimated prevalence from influenza-related Internet searches, drastically overestimated peak flu levels compared to traditional surveillance data, leading Google to finally close

their website in August 2015. Other experiments, still on influenza, relating more generally to syndromic surveillance, were carried out based on the analysis of social networks, for example, Twitter, with apparently better results, which remain to be replicated and confirmed [20]. These methods can thus complement, but not replace, traditional epidemiological surveillance networks [5].

The unprecedented development of connected health tools and applications already offers the possibility for those who benefit from these tools to know and monitor certain elements of their health status (weight, blood pressure, heart rate, glycemia, calories ingested and expended, etc.), and therefore in theory to be able to detect abnormalities and disturbances before the disease sets in. We can anticipate that the development of this type of tools will increase in the next years, allowing us to know more and more elements of our health. The already preponderant role of this type of tools in health surveillance was highlighted by the COVID-19 epidemic for which several applications have been developed to know its status with regard to infection and exposure to contact cases.

The question of the validity of these tools to measure what they are supposed to measure, of their access for all, of the storage and secondary use of the information coming from these tools most of the time developed by private companies is a major issue.

Pharmacovigilance

Pharmacovigilance may be defined as the collection of information useful for monitoring drugs, including information on suspected adverse reactions when a drug is used. It is classically based on the careful analysis of reports by doctors to public health authorities and/or pharmaceutical companies. In normal practice, it takes time after a medication has been put on the market before the medication is found to have caused an undesirable event. The current pandemic underscores the need and importance of accurate, up-to-date information to enable us to judge the safety and risk of vaccination against

COVID-19 almost in real time. The use of a variety of databases, from electronic medical records to medico-administrative databases and Internet, can constitute an opportunity to achieve this goal and to identify potential side effects more quickly.

Previous studies have already shown the ability of using AI approaches to detect evidence of unreported prescription drug side effects. For example, White R et al. were able to find evidence, by using automated software tools to examine queries handled by six million Internet users taken from Web search logs in 2010, that the combination of an antidepressant, paroxetine, and a cholesterol lowering drug, pravastatin, caused high blood sugar [21].

The application of such methods requires extensive expertise in computer science and statistics that was not required by standard adverse event monitoring. It is difficult to imagine that these methods and associated tools are not destined to be used more and more in the future development of pharmacovigilance. However, these tools will not replace human pharmacovigilance, but they could be of considerable help.

Sentiment Analysis and the Use of Social Media as Alternative Data Source and Data Processing Method in Epidemiology

Virtually, the development of AI makes it possible to process any form of digital data, especially if it comes in large volumes. A new use in epidemiology is therefore to take advantage of new sources of data: in theory, any data from the Web, in practice, data from forums or, more generally, from social media. These networks can be used and their content analyzed in different ways: They can allow the active or passive recruitment of people for the constitution of e-cohorts, or syndromic surveillance.

Another approach to using AI in epidemiology is the analysis of content combined with volume and uses linked to social media. In particular, sentiment analysis is increasingly used.

Recruitment of E-cohorts and Online Syndromic Surveillance

Faced with the cost and the necessary logistics linked to the constitution and then to the maintenance of traditional cohorts, which are increasingly large in number of people, and also because of the evolution of the means of access to people and uses, alternative strategies or complementary products are developing in epidemiology, aiming both to reach and maintain as many people as possible in the cohorts, and to reduce maintenance costs. Electronic cohorts, or e-cohorts, have therefore developed, to complement or replace, for example, cohorts based on the sending of paper questionnaires or face-to-face evaluations. Of course, the profile of the populations recruited by traditional approaches and by “2.0” approaches, as well as the way of answering the questions (and of asking them), cannot be strictly superimposed.

It is possible to go further than substituting a paper form for a Web form, for example, by recruiting people directly via social networks. Facebook is one such example [22]. Either actively: creation of a page and call to participate via the media; or passively, by analysis and selection of content and profiles obtained by data collection (for example, on patient forums).

Social Media Content and Sentiment Analysis

Textual data, especially in large volumes, are particularly suitable for analysis by AI, in the sense that a certain amount of information, the quality and relevance of which must be carefully considered, can be extracted automatically where a human operator would not have enough of a life to read everything and then synthesize. Text mining therefore consists of the analysis of textual data, from the simplest to the most sophisticated. A particular use of text mining and social media is sentiment analysis. It can be defined as follows: determine whether the feeling given off by a sentence is positive or negative [23]. The main difficulty of analysis lies at the very heart of the use of language. Among other factors, the feeling of a sentence depends on context and language, as well as on the person who wrote it. Sentiment

analysis can be based on two approaches: lexical analysis and analysis based on machine learning. In both cases, AI is strongly associated with it. Both methods can be combined [24]. Lexical analysis will use semantic analysis, based on a classification of the sentence studied against dictionaries, for which words or sentences deemed to be similar have been annotated in terms of negative or positive polarity.

Several examples of social media analysis can be mentioned. For example, in order to better understand the representations and behaviors vis-à-vis vaccination. Authors used Twitter to update the semantic web of positive, negative, and neutral feelings about vaccines [25]. Beyond this static representation, it is possible to study the relationships between getting vaccinated and epidemic dynamics (Disneyland measles outbreak 2015 in California, in [26]). The results can be used to guide public health policies and adapt vaccination campaigns and communication around vaccination according to the representations and behaviors of populations. Another similar example concerns the Middle East Respiratory Syndrome (MERS) outbreak in Korea in 2015 [27]. Epidemiological indicators can therefore be derived from these analyses of social networks. However, like any indicator, these “digital” indicators require rigorous validation [28].

Conversely, social networks can be used as a communication channel for the dissemination of public health messages [29].

From Data to Decision: Data-and Model-Driven Knowledge and Decision in Epidemiology

The emergence of AI in any field always raises the question of what task will be automated, and therefore, to what extent AI will act as a substitute for humans. An optimistic view of the use of AI resonates with that of industrialization, then of automation and robotization of assembly lines in factories, and which would be that AI would free humans from tasks “unworthy” of his condition and would allow him to concentrate either on

tasks specific to human skills or on his hobbies. In epidemiology, then in its possible uses in public health, the place of the human being intervenes at all stages, from the collection of data to the interpretation of the analyses and the results, possibly the decision-making, passing by technical choices in terms of data and problem modeling.

We have seen that AI can intervene from the stages of data collection, reconciliation, and then structuring, in particular under conditions that would be difficult to access by humans in a reasonable time and resources. It can intervene at the data analysis stage, as complementary techniques to the more traditional ones used until then. It can still be placed a little further in the value-added chain of data processing: in the more or less automated choice of the “best” models (and “best” predictors), under the assumption that data-driven approaches would be more atheoretical than the choice of a model by a human operator – which remains a hypothesis at least naïve; in the automation of monitoring an entire part of the health care system and the health status of the population, triggering alarms that will require the attention of a human operator to decide whether the alarm is relevant or not; and finally, in helping, and even in decision-making in terms of public health management, directly based on the use of collected data.

The first use case is similar to 4P medicine (predictive, personalized, participatory, and preventive). We can, for example, mention the data-driven approaches to predict the occurrence of diabetic and cardiovascular diseases [30]. Some authors highlight the inability of human operators to effectively mix and synthesize large datasets, for example, data from the Global Burden of Disease study. AI is then used there to screen all the data and draw attention to potentially interesting results [31].

More ambitious, some projects and systems promise to make projections over several decades in terms of needs and use of the health system. Authors have produced such a 40-year projection model for Singapore, in order to take into account the impact of different factors such as ethnicity, social isolation, or disability [32]. Still other authors have endeavored to develop and then

calibrate an AI that they called automated time series machine learning, applied to Romanian data and making it possible to predict the health of the population over the coming years in relation to the ten most common diseases, more deadly [32].

Finally, to the extreme, some authors have taken the reasoning to the limit, sweeping aside any need for causal and explanatory models of diseases and health events, for any use of statistical inference and the setting up of controlled experiments: Data would be everywhere, would characterize everything, and its increasing accessibility coupled with the growing efficiency of AI would be enough that for any question and need for a decision, there would be an AI to answer it. This would be the “end of the theory” – here, the end of epidemiology, in its exploratory dimension and consisting in investigating the causes of the states of health of populations [33]. This vision is obviously highly contested [34].

Potential Trends and Future Challenges

The Challenges Related to Data

Since its first successes, epidemiology has still faced many challenges, in the extension of its first missions and following the natural evolution of its field of study, according to discoveries and the improvement of study methods. In particular, it is no longer possible to remain focused on an archetypal theoretical model restricted to “an exposure, a cause, a disease.” The challenges of modern epidemiology include the identification and simultaneous consideration of multiple risk factors, and also the long-term effects, whether low or high intensity, direct or mediated or resulting from interactions, or finally the notion of “personalized” medicine. The availability and possibility of combining data increasingly varied in nature and sources, the real and current possibility of secondary use of data (for example, that of hospital health data warehouses), seems particularly favorable the use of AI in epidemiology to keep it developing. However, controlling the quality of data is an imperative in epidemiology

– in reality, controlling the entire data chain: from its collection to the interpretation of analyses. The reasonable use of new data sources will have to go through the exploration and characterization of their quality: data from connected objects, behavioral data, data from social networks, etc. It is illusory to be able to directly derive robust epidemiological knowledge from the blind collection of unqualified data. In particular, the increasing secondary use of data raises the question of value of the data, the purpose for which the data has been originally produced [35]. The objective, method, and quality that guide data collection shape the information that is produced.

The potential contribution of big data and AI appears significant insofar as information on a wide range of characteristics of the environment or context of life (social, economic, and cultural) can be collected and be connected with health data, for example, to develop models on social determinants of health including individual lifestyle factors, social and community networks, and general socio-economic, cultural, and environmental conditions. The underlying assumption is that in the field of health, data from several bases, with their diversity, are likely to generate information and knowledge on health and the environment which builds it. Such approach of health is relying increasingly on the use of data not specifically collected for that purpose, including data *a priori* not related to health. Interestingly, this potentially broadened approach to health research that may be facilitated by big data and AI is largely underused in practice. The notion of big data and AI remains often focused on the use of large volumes of data, often biological (genome, omic), for individual purposes. Thus, it is aimed at predictive medicine, the determination of individual risks or diagnostic decisions through the multiplication of individual biological data, for an “*a la carte*” health, popularized by the term “personalized medicine” [35]. Far from allowing us to develop a “personified” medicine [36], integrating biological, clinical, psychosocial, and environmental elements to apprehend more broadly the determinants of health of an individual [35], the development of big data in health and personalized medicine tend to lead to a reductionism of the individual at its biological dimension.

The Challenges Related to Epidemiology as an Explanatory Science: Statistical Inference and Causality

C Andersen announced “the end of the theory” in 2008 in Wired, amid the explosion of data and data analysis capabilities [33]. This amounts to saying that the explanatory dimension of science is no longer a necessity, that correlation is sufficient and that prediction alone can and must guide our individual or collective actions. The establishment of knowledge making history and meaning for humans would be obsolete. In fact, all efforts on causal inference, the establishment of more or less complex causal links between given exposures and outcomes, are canceled. At the same time and following on from the work that occupied most of J Pearl’s life on the statistical approach to causality [37], several techniques relating to AI are progressing on the path to the discovery of interrelations, sometimes causal under certain well-defined assumptions as to the nature of the data used, from masses of data. This is the case with the numerous structure learning and learning algorithms of Bayesian networks [38]. Curiously, these methods have so far not interested epidemiologists much, even though they are applied more and more frequently in other fields, including those of industrial risk control. Between these two approaches, the one which denies the need for causality and the one which claims to automatically discover the causal networks in the masses of data, some authors explain that for certain situations, causality may on the one hand be unattainable, and secondly not necessarily useful. In these situations, in particular in the case of health policy and management of the care system, robust and precise prediction could be a tool synonymous with progress [39]. Nevertheless, there are cases where causality seems inevitable: In clinical epidemiology, the cause is sought in order to be able to act on it, and therefore prevent or treat people.

However, although the question of why cannot be essential for commercial uses or marketing, as opposed to the question of what, but it is essential in the specific field of health and intervention in

this area. Trying to identify factors to improve health without the prior attempt to approach the causes cannot be a relevant and efficient approach to health. At a time when the volume of data – both structured and unstructured – becomes difficult to use with conventional solutions, it becomes crucial to achieve using data from various sources and of various types to produce a “different” information certainly, and also information that is more complete and more reliable. The search for causal and veracity of data are more than ever of the major challenges in terms of big data, even in the specific field of health [40].

The Challenges Related to the Use Made of Epidemiology by Decision-Makers and the Involvement of Private Actors

In the end, it must be said that epidemiology, what it produces, to be useful, does not remain in the sole hands of epidemiologists. First, the actors involved in epidemiology and public health have changed somewhat over the past 20 years, not least because of the importance of data and the growing interest in new data sources. Thus, all private platforms, Google, Amazon, Facebook, and Apple (GAFAMS) have at one time or another, at one stage or another in the data processing chain, ambitions in the field of individual and population health. Their fundamental role as intermediaries leads them to position themselves not only as intermediaries but also as service provider due to their capacity to capture and process data. Finally, at the end of the data processing chain is the decision-maker, usually a policymakers. The experience of the COVID-19 pandemic underlines how the decision often remains above all political and unscientific approaches, for various reasons – the logistical reason not being the least (stocks of masks, number of hospital beds, availability of vaccines, etc.). In this context, the value of the epidemiologist as a researcher and human operator can become uncertain, and the use of AI both as a technical tool and as a guarantee for policy makers a real opportunity. This use of AI by policy makers,

without understanding how these results are produced and their limits, is a significant risk for health decision-making. The example of decisions taken in the face of the COVID-19 epidemic on the basis of partial data, imperfect models, is an illustration of this.

Conclusion

Since its origin, epidemiology has always been built on data and its interpretation, which makes it a discipline particularly concerned by the development of massive data and AI. Numerous uses are already in progress and are likely to develop. This development is not without raising major challenges for epidemiology in its ability to describe health of the whole populations, and identify risk factors accessible for action, while ensuring the quality and veracity of data. Over the course of its history, epidemiology has already had to deal with the integration of new data sources and new methods of analysis. We can therefore be confident in its ability to best integrate the exploitation of massive data and AI methods, remembering that these are methods and tools with evident advantages but also limitations, but that they remain tools, and not the solution, to better understand health.

References

1. Bonita R, Beaglehole R, Kjellstrom T. Elements of epidemiology (French); 2010.
2. Last JM. A dictionary of epidemiology. 4th ed. Oxford: Oxford University Press; 2001.
3. Valleron J. Les rôles de la modélisation en épidémiologie. Comptes Rendus de l'Académie des Sciences - Series III - Sciences de la Vie. 2000;323(5):429–33.
4. Laney D. 3D data management: controlling data volume, velocity, and variety. Rome: Application Delivery Strategies Meta Group; 2001.
5. Flahault A, Bar-Hen A, Paragios N. Public health and epidemiology informatics. Yearb Med Inform. 2016;1: 240–6. <https://doi.org/10.15265/IY-2016-021>.
6. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA. 2018;319(13):1317–8. <https://doi.org/10.1001/jama.2017.18391>.
7. Huang M, Gibson C, Terry A. Measuring electronic health record use in primary care: a scoping review. Appl Clin Inform. 2018;9(1):15–33.

8. Greenhalgh T, Hinder S, Stramer K, Bratan T, Russell J. Adoption, non-adoption, and abandonment of a personal electronic health record: case study of HealthSpace. *BMJ*. 2010;341:c5814.
9. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019;380(14):1347–58.
10. Thiébaut R, Thiessard F. Section editors for the IMIA yearbook section on public health and epidemiology informatics. Artificial intelligence in public health and epidemiology. *Yearb Med Inform*. 2018;27(1):207–10. <https://doi.org/10.1055/s-0038-1667082>.
11. Unberath P, Prokosch HU, Gründner J, Erpenbeck M, Maier C, Christoph J. EHR-independent predictive decision support architecture based on OMOP. *Appl Clin Inform*. 2020;11(3):399–404. <https://doi.org/10.1055/s-0040-1710393>.
12. Choudhury O, Park Y, Salomidis T, Gkoulalas-Divanis A, Sylla I, Das AK. Predicting adverse drug reactions on distributed health data using federated learning. In: AMIA annual symposium proceedings, 2020 Mar 4; 2019. p. 313–322.
13. Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated learning for healthcare informatics. *J Healthc Inform Res*. 2020:1–19. <https://doi.org/10.1007/s41666-020-00082-4>.
14. Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci Rep*. 2020;10(1):12598. <https://doi.org/10.1038/s41598-020-69250-1>.
15. Luce BR, Connor JT, Broglio KR, Mullins CD, Ishak KJ, Saunders E, Davis BR, RE-ADAPT (REsearch in ADaptive methods for Pragmatic Trials) Investigators. Using Bayesian Adaptive trial designs for comparative effectiveness research: a virtual trial execution. *Ann Intern Med*. 2016;165(6):431–8.
16. Dolgin E. Industry embraces virtual trial platforms. *Nat Rev Drug Discov*. 2018;17(5):305–6. <https://doi.org/10.1038/nrd.2018.66>.
17. Thacker SB, Berkelman RL. Public health surveillance in the United States. *Epidemiol Rev*. 1988;10:164–90.
18. Chiolero A, Buckeridge D. Glossary for public health surveillance in the age of data science. *J Epidemiol Community Health*. 2020;74:612–6.
19. Ginsberg J, Mohebbi M, Patel R, Brammer L, Smolinski M, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature*. 2009;457:1012–4.
20. Huang P, MacKinlay A, Yepes AJ. Syndromic surveillance using generic medical entities on Twitter. In: Proceedings of Australasian Language Technology Association Workshop; 2016. p. 35–44.
21. White R, Tatonetti N, Shah N, Altman R, Horvitz E. Web-scale pharmacovigilance: listening to signals from the crowd. *J Am Med Inform Assoc*. 2013;20(3):404–8.
22. Fenner Y, Garland SM, Moore EE, Jayasinghe Y, Fletcher A, Tabrizi SN, Gunasekaran B, Wark JD. Web-based recruiting for health research using a social networking site: an exploratory study. *J Med Internet Res*. 2012;14(1):e20. <https://doi.org/10.2196/jmir.1978>.
23. Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM. Temporal patterns of happiness and information in a global social network: hedonometrics and twitter. *PLoS One*. 2011;6(12):e26752. <https://doi.org/10.1371/journal.pone.0026752>.
24. Mudinas A, Zhang D. Levene combining lexicon and learning based approaches for concept-level sentiment analysis. In: Proceedings of the first international workshop on issues of sentiment discovery and opinion mining (WISDOM '12), vol. 5. New York: Association for Computing Machinery; 2012. p. 1–8. <https://doi.org/10.1145/2346676.2346681>.
25. Kang GJ, Ewing-Nelson SR, Mackey L, Schlitt JT, Marathe A, Abbas KM, et al. Semantic network analysis of vaccine sentiment in online social media. *Vaccine*. 2017;35:3621–38.
26. Pananos AD, Bury TM, Wang C, Schonfeld J, Mohanty SP, Nyhan B, et al. Critical dynamics in population vaccinating behavior. *Proc Natl Acad Sci*. 2017;114:201704093.
27. Choi S, Lee J, Kang MG, Min H, Chang YS, Yoon S. Large-scale machine learning of media outlets for understanding public reactions to nation-wide viral infection outbreaks. *Methods Inf Med*. 2017;129:50–9.
28. Margulis AV, Fortuny J, Kaye JA, Calingaert B, Reynolds M, Plana E, et al. Value of free-text comments for validating cancer cases using primary-care data in the UK. *Epidemiology*. 2018;29:308–13.
29. Gough A, Hunter RF, Ajao O, Jurek A, McKeown G, Hong J, et al. Tweet for behavior change: using social media for the dissemination of public health messages. *JMIR Public Health Surveill*. 2017;3:e14.
30. Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Med Inform Decis Mak*. 2019;19(1):211. <https://doi.org/10.1186/s12911-019-0918-5>.
31. Flaxman AD, Vos T. Machine learning in population health: opportunities and threats. *PLoS Med*. 2018;15(11):e1002702. <https://doi.org/10.1371/journal.pmed.1002702>.
32. Chan CL, Chang CC. Big data, decision models, and public health. *Int J Environ Res Public Health*. 2020;17(18):6723. <https://doi.org/10.3390/ijerph17186723>.
33. Anderson C. The end of theory: the data deluge makes the scientific method obsolete. *Wired*; 2008. <https://www.wired.com/2008/06/pb-theory>
34. Pigliucci M. The end of theory in science? *EMBO Rep*. 2009;10(6):534. <https://doi.org/10.1038/embor.2009.111>.
35. Delpierre C, Kelly-Irving M. Big data and the study of social inequalities in health: expectations and issues. *Front Public Health*. 2018;6:312.
36. Ziegelstein RC. Personomics. *JAMA Intern Med*. 2015;175(6):888–9. <https://doi.org/10.1001/jamainternmed.2015.0861>.

37. Pearl J. An introduction to causal inference. *Int J Biostat.* 2010;6(2):7. <https://doi.org/10.2202/1557-4679.1203>.
38. Lefèvre T, Lepresle A, Chariot P. Detangling complex relationships in forensic data: principles and use of causal networks and their application to clinical forensic science. *Int J Legal Med.* 2015;129(5):1163–72. <https://doi.org/10.1007/s00414-015-1164-8>.
39. Kleinberg J, Ludwig J, Mullainathan S, Obermeyer Z. Prediction policy problems. *Am Econ Rev.* 2015;105(5):491–5.
40. Dimeglio C, Kelly-Irving M, Lang T, Delpierre C. Expectations and boundaries for big data approaches in social medicine. *J Forensic Legal Med.* 2018;57:51–4. <https://doi.org/10.1016/j.jflm.2016.11.003>.



Cécile Nabet, Aniss Acherar, Antoine Huguenin, Xavier Tannier,
and Renaud Piarroux

Contents

Introduction	1354
Malaria Diagnosis: Parasite Detection and Species Identification	1354
Vector Ecology: Species, Biology, and Behaviors	1355
Deployment of Artificial Intelligence in Malaria	1356
Applications of Artificial Intelligence in Malaria Diagnosis and Malaria Vector Characterization	1357
Image-Based Automatic Classification for Malaria Diagnosis	1357
Image-Based Automatic Classification of Mosquito Vectors	1361
Characterization of <i>Anopheles</i> Biological Features Using Proteomic Tools	1364
Conclusion	1366
Cross-References	1366
References	1366

Abstract

Malaria disease is due to the infection with *Plasmodium* parasites transmitted by a mosquito vector belonging to the genus *Anopheles*. To combat malaria, effective diagnosis and treatment using artemisinin-based combinations are needed, as well as strategies that are aimed at reducing or stopping transmission by mosquito vectors. Even if the conventional microscopic diagnosis is the gold standard for malaria diagnosis, it is time consuming, and the diagnostic performance depends on techniques and human expertise. In addition, tools for characterizing *Anopheles* vectors are limited and difficult to establish in the field. The advent of computational biology, information technology infrastructures, and mobile

C. Nabet (✉) · A. Acherar · R. Piarroux
Sorbonne Université, Inserm, Institut Pierre-Louis
d’Epidémiologie et de Santé Publique, IPLESP, AP-HP,
Groupe Hospitalier Pitié-Salpêtrière, Service de
Parasitologie-Mycologie, Paris, France
e-mail: cecilenabet7@gmail.com;
aniss.acherar@gmail.com; renaud.piarroux@aphp.fr

A. Huguenin
EA 7510, ESCAPE, Laboratoire de Parasitologie-
Mycologie, Université de Reims Champagne-Ardenne,
Reims, France
e-mail: ahuguenin@chu-reims.fr

X. Tannier
Sorbonne Université, Inserm, Université Sorbonne Paris
Nord, Laboratoire d’Informatique Médicale et d’Ingénierie
des Connaissances pour la e-Santé, LIMICS, Paris, France
e-mail: [xavier.tannier@sorbonne-universite.fr](mailto:xavier.tannier@ sorbonne-universite.fr)

computing power offers the opportunity to use artificial intelligence (AI) approaches to address challenges and technical needs specific to malaria-endemic countries. This chapter illustrates the trends, advances, and future challenges linked to the deployment of AI in malaria. Two innovative AI approaches are described. The first is the image-based automatic classification of malaria parasites and vectors, and the second is the proteomics analysis of vectors. The developed applications are aimed at facilitating malaria diagnosis by performing malaria parasite detection, species identification, and estimation of parasitaemia. In the future, they can lead to efficient and accurate diagnostic tools, revolutionizing the urgent diagnosis of malaria. Other applications focus on the characterization of mosquito vectors by performing species identification, behavior, and biology descriptions. If field-validated, these promising approaches will facilitate the epidemiological monitoring of malaria vectors and saving resources by preventing or reducing malaria transmission.

Keywords

Malaria · *Anopheles* · *Plasmodium* · Vector control · Malaria diagnosis · Artificial intelligence · Machine learning · Artificial neural networks · MALDI-TOF MS

Introduction

Malaria is a disease caused by infection with *Plasmodium* parasites and transmitted by a mosquito vector belonging to the genus *Anopheles* [1]. These pathogens are transmitted to humans during the blood meal of infected female *Anopheles*. *Plasmodium* proliferation in red blood cells is responsible for chills and fever, the main symptoms of the disease [2]. Five species are involved in human malaria: *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium ovale*, *Plasmodium malariae*, and *Plasmodium knowlesi*. The most common and deadliest species is *P. falciparum*. Despite global malaria control efforts, the disease

persists, and approximately 229 million cases and 409,000 deaths were estimated to have occurred globally in 2019 by the World Health Organization [3]. The most primarily affected people are the sub-Saharan African population and children under 5 years of age. To combat malaria, direct and indirect control strategies have been developed. Direct strategies are aimed at curing malaria cases thanks to effective diagnosis and treatment using artemisinin-based combinations. Indirect strategies are aimed at reducing or stopping transmission by mosquito vectors through adapted vector control programs [4, 5].

Malaria Diagnosis: Parasite Detection and Species Identification

Malaria is a major cause of febrile illness in endemic countries. It requires rapid therapeutic management, especially due to the critical lethality of *P. falciparum* in non-immune populations [2]. However, such a clinical presentation is unspecific, and other causes of severe febrile illnesses are increasingly described [6]. Therefore, parasitological confirmation of malaria is necessary to guide healthcare workers in the management of patients with fever. Microscopic diagnosis by staining thick and thin blood smears is the most commonly used method for malaria diagnosis [2, 7]. Thick blood smears concentrate malaria parasites and are used for screening to detect the presence of *Plasmodium* parasites. A well-trained malaria microscopist should be able to recognize the *Plasmodium* species correctly in thick blood smears, even at relatively low parasite density. Indeed, the limit of detection of Giemsa-stained thick blood smears by expert malaria microscopists is estimated to be between 4 and 20 parasites per μl [7]. Thin blood smears are used for morphological differential diagnosis of *Plasmodium* species, analyzing erythrocyte details such as size, shape and crenation, characteristic dots in the erythrocyte stroma, parasite pigment structure and color, and parasite stages such as trophozoites (including ring stages), schizonts, and gametocytes. Nevertheless, when Giemsa staining contains very few parasites, differential

diagnosis is more difficult, and errors in species diagnosis are more common [7]. Thin blood smears are also used to estimate the parasite density of *P. falciparum*. For quantification using the thin film, many successive fields (a minimum of 100 fields at x100 magnification, which is equivalent to at least 2,000 erythrocytes) are counted for infected red blood cells (RBCs). Parasite density (also called parasitemia) is expressed according to the equation $\% \text{Parasitemia} = \frac{\text{Parasitized RBCs}}{\text{Total RBCs}} \times 100$ [8]. Fields used to estimate parasitemia should contain little to no overlap of red blood cells. When parasitized RBCs are counted, multiple infected RBCs should be counted only once. Gametocytes are not counted, as they are a dead-end stage in the human host. Parasite density is indicative of the severity of the disease, and mortality of *P. falciparum* malaria rises when the proportion of infected erythrocytes increases, although the relation between parasite density and prognosis is variable [2]. Examination of serial blood smears with parasite density quantification is also recommended for monitoring response to therapy.

Even if conventional microscopic diagnosis is the gold standard for malaria diagnosis, it is time consuming, and the diagnostic performances depend on techniques and human expertise. This expertise is rare, not only in malaria-endemic countries due to the lack of adequate initial training and scarcity of quality control but also in malaria eliminating countries and in non-endemic countries due to the lack of experienced microscopists. Indeed, routine microscopy diagnostics have been shown to be inaccurate in many health care centers across Africa [9–11], with a high rate of false positive results as high as 75.6% [11], leading to a frequent over-diagnosis of malaria. In addition, high variability in the estimation of parasite density is encountered due to the use of various methods [7]. Alternatives have been developed for rapid diagnosis in the field in the absence of microscopy infrastructures and experts, such as malaria rapid diagnostic tests (RDTs), to detect *Plasmodium* antigens. The accuracy of *Pf*HRP2-based tests for *P. falciparum* diagnosis is equivalent to that of routine microscopy, and they can be used as a screening tool in

resource-limited settings [2]. However, the identification of non-*P. falciparum* species is not feasible using each malaria RDT, and/or the sensitivity for these species is lower [7, 8]. In addition, malaria RDTs do not inform the parasitemia or the parasite stages, two crucial pieces of information for optimal clinical malaria case management.

Vector Ecology: Species, Biology, and Behaviors

Malaria vector control measures such as insecticide-impregnated bednets, insecticide spraying, or bacterial toxins are widely used to kill adult mosquitoes or larvae [5]. These measures are essential for controlling mosquito transmission of malaria parasites. However, to deploy effective malaria vector control measures, knowledge of *Anopheles* vector ecology is crucial. *Anopheles* mosquitoes are dipterous insects that develop in two phases: an aquatic phase for immature life stages and an air phase for adults. Adult *Anopheles* have a body segmented into three anatomic parts (head, thorax, and abdomen) and hosts three pairs of legs and a pair of wings attached to the thorax for locomotion. Only some *Anopheles* species are true vectors of malaria. Thus, the first step consists of identifying the *Anopheles* species. Species identification is usually performed using dichotomous keys based on the morphologic characteristics of collected specimens. It requires highly technical skills acquired after comprehensive training. For instance, morphometry of wing venation characteristics is a tool used by experts to discriminate between genera and species of mosquitoes. An additional challenge is identifying cryptic *Anopheles* species belonging to the same complex of species, as they do not have the same capacity to transmit *Plasmodium*. However, they are morphologically identical and can only be distinguished based on molecular tools, an alternative to morphology methods to identify *Anopheles* vector species [12]. Such tools are efficient, but the time required for sample preparation and the subsequent costs limit their use for extensive screening.

Protein profiling using matrix-assisted laser desorption ionization-time of flight (MALDI-

TOF) mass spectrometry (MS) is a promising new tool. Generated mass spectra protein profiles are fingerprints and characteristics of an organism (species, strain, or physiological stage). The method is rapid and accurate and is suitable for routine applications. Thanks to a comparison with a reference mass spectra database, it has been widely used in clinical microbiology for the species identification of bacteria [13], fungi [14], and parasites [15]. More recently, it has been developed for arthropod vector identification [16]. Several teams have built in-house databases to identify species of adult *Anopheles* by their MALDI-TOF spectra [17]. If deployed for mosquito surveillance, this proteomic method will avoid molecular analysis, with a gain of rapidity and cost of analysis. However, further research is needed to improve the resolution for cryptic species using new bioinformatic tools and for better standardization. Similar to DNA sequence databases, accessibility through online applications is essential for the broad use of MALDI-TOF MS databases. Such online platforms have already been proposed for fungi [14] and *Leishmania* species [18] and are currently being set up for mosquito species identification [17].

Aside from species identification, there is a need to perform more complex tasks to study *Anopheles* biology patterns, such as *Plasmodium* infection, past blood meal, age, and mosquito behaviors. These determinants of malaria transmission are particularly useful for vector monitoring. Evidence of *Anopheles* infection by *Plasmodium* is a prerequisite to confirm the role of a given species as a vector, and the proportion of individuals hosting sporozoites in their salivary glands (thorax) is a determinant of the capacity of malaria transmission [1]. Confirmation of a previous blood meal in captured specimens provides information about the human biting behavior of the population of mosquitoes. The human biting rate helps quantify malaria transmission. Within a given population of mosquitoes, only a small proportion of specimens can transmit malaria. Transmission only occurs after an infected blood meal, and individuals become infectious to humans only until parasite incubation is achieved (approximately 9–14 days, sometimes longer depending on the species and the temperature conditions). Thus, the average age of

the *Anopheles* female mosquito population is also an important determinant of the likelihood of malaria transmission since only the oldest mosquitoes in a population are responsible for *Plasmodium* transmission. Finally, exploring *Anopheles* mosquito behavioral patterns, such as blood feeding, is useful for understanding behaviors relevant to parasite transmission. All of these *Anopheles* mosquito characteristics may be useful for implementing adapted vector control measures and improving efficacy monitoring.

Overall, tools for characterizing *Anopheles* vectors are limited and difficult to implement in the field [19]. Apart from PCR tools that can detect *Plasmodium* in mosquitoes, methods are mainly based on laborious microscopy examination and require fresh material and technical skills [1]. In particular, there is a need for operationally attractive methods to characterize the *Anopheles* vectors, especially age, which is not accessible by DNA analysis. In addition, mosquitoes' feeding behaviors remain poorly understood, especially as they require scarifying researcher or mouse skin.

Deployment of Artificial Intelligence in Malaria

Technology now makes it possible to take photographs, including from a microscope, or to generate a mass spectrum from a biological sample and send images and spectra to an internet platform. These possibilities open the field to diagnostic procedures that can be activated from almost anywhere in the world. To finalize an online diagnostic system, it is necessary to make available automatic analysis methods of images and spectra obtained in the field. This opens up an avenue for artificial intelligence (AI) approaches addressing challenges and technical needs that are specific to the field of malaria-endemic countries, bringing new tools for malaria control [20, 21]. Progress in computational biology along with improved access to information technology infrastructures and mobile computing power in many low- and middle-income countries offers the opportunity to deploy AI-driven health interventions [22]. AI involves supervised machine learning methods, which are a set of algorithms applied to already labeled data to learn a statistical

model for pattern recognition, classification, or prediction. Such models must be able to generalize the learned task to new, unseen data. Most of them rely on artificial neural networks (ANNs), a class of machine learning algorithms. Among them, deep ANNs such as convolutional neural networks (CNNs) are specific architectures of deep learning models able to produce a reduced representation from sequences of elements (e.g., images as a sequence of pixels, text as a sequence of words) that are particularly well suited for image-based automated classification. There are various methods to train a neural network, one of them being to adapt pretrained models and use transfer learning approaches. Transfer learning consists of using the knowledge acquired on a general classification problem to apply it again to a particular classification problem.

In the following paragraphs, the trends, advances and future challenges linked to the deployment of AI in malaria are illustrated. Two innovative AI approaches are described in this chapter. The first is the image-based automatic classification of malaria parasites and vectors, and the second is the proteomics analysis of vectors. The developed applications are aimed at facilitating malaria diagnosis by performing malaria parasite detection, species identification, and estimation of parasitaemia. Other applications focus on characterizing mosquito vectors by performing species identification, behavior, and biology descriptions. Methods involving AI to model malaria transmission and vector ecology are considered indirect applications and will not be discussed in this chapter.

Applications of Artificial Intelligence in Malaria Diagnosis and Malaria Vector Characterization

Image-Based Automatic Classification for Malaria Diagnosis

For automated microscopy malaria diagnosis, three elements are crucial:

- Sample preparation, including smear realization and staining

- Acquisition of digital images using a microscope
- “Computer vision” algorithm to analyze and classify the captured digital images

Digital images are a finite set of digital values, called “picture elements” or “pixels.” The digital image contains a fixed number of rows and columns of pixels. Pixels are the smallest individual element in an image, holding values that represent the brightness of a given color at any specific point. Usually, a digital image is mathematically processed using the red green and blue (RGB) representation that converts a pixelated image into a matrix of intensity values of red, green, and blue (Fig. 1).

An automated image-based classification system for malaria diagnosis is a multistep process composed of two main processes: data processing and classification (Fig. 2). Data processing includes image acquisition and preprocessing techniques such as noise reduction and normalization. The classification of digital images consists of counting the infected and non-infected red blood cells (RBCs) and classifying *Plasmodium* according to its species and life stages. Traditional approaches to detecting malaria-causing parasites on thin smears involve specific data processing steps, such as feature engineering. Feature engineering consists of extracting patterns and signatures from images that can help identify specific objects such as RBCs (also called regions of interest). Prominent features are then selected to allow classification of infected and non-infected RBCs, such as morphology (shape and texture information), color, and granulometry. Different color space representations can be used, such as the hue saturation value (HSV) [23, 24] or the L*a*b* color, where L* represents the lightness, a* characterizes the red/green value, and b* characterizes the blue/yellow value, which is the more prominent color channel to characterize malarial parasites [25]. Traditional approaches use conventional machine learning models, such as multi-class support vector machines (SVMs) [26], naïve Bayes, or moving K-means [23, 24, 27]. Overall, the classification performance of such machine learning models ranged from 80% to 99% accuracy for *Plasmodium* species identification. However,

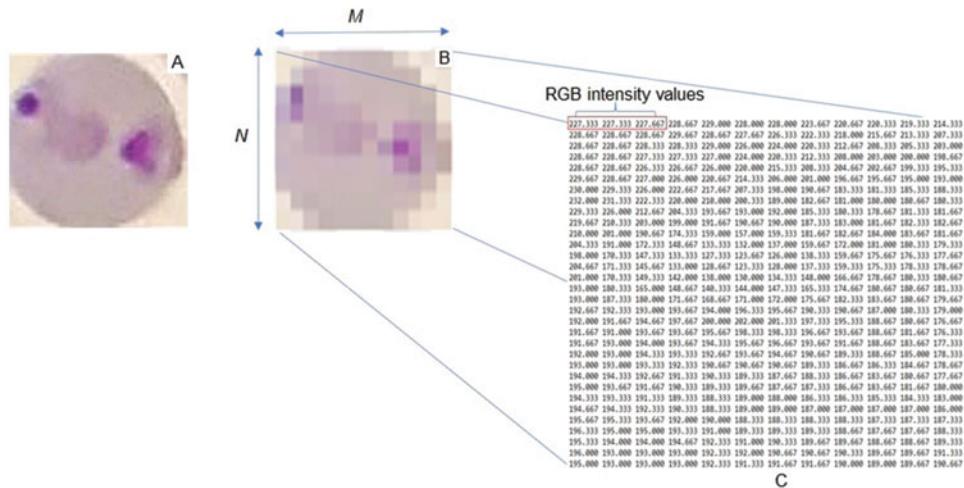


Fig. 1 RGB image representation: (a) *Plasmodium falciparum* infected red blood cell. (b) Pixelated image with the shape ($M \times N$) where M is the number of columns and N the number of rows. (c) The matrix of corresponding

intensity values defined by three values, red, green, and blue (RGB) in a range of 256 possible values: a mixture of these three values produces color in the image

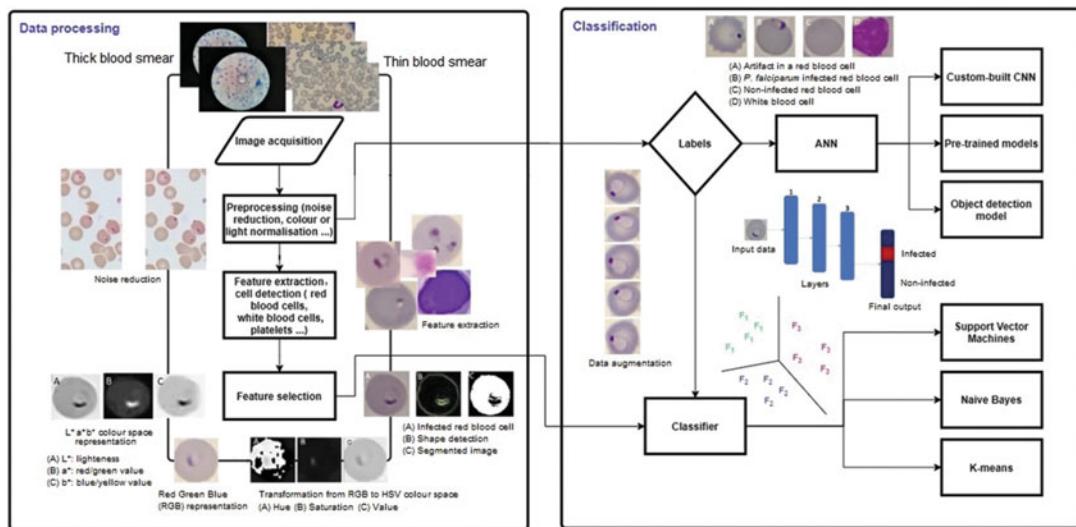


Fig. 2 Key steps of an automated image classification system for malaria diagnosis

evaluations were set on a relatively small number of images, corresponding to a limited number of blood smears (approximately one hundred). Most of these studies have proposed oversimplified solutions based on unicentric datasets, hence not applicable for routine diagnosis due to their low generalization capacity and their lack of robustness. The variety of staining methods and quality of smear preparation for microscopic blood smear

images makes it difficult to devise universal features using traditional approaches. Finally, a manual design of the features of interest requires expertise and is laborious and time consuming, especially for discriminating *Plasmodium* species and life stages.

With the advent of deep learning, ANNs, especially convolutional neural networks (CNNs), have achieved expert-level performance in the

detection of many pathologic characteristics. The high-performances mainly rely on medical image analysis, such as scans for retinal disease diagnosis [28] or histopathology slides for cancer disease diagnosis [29], to cite a few examples. Unlike feature computation methods, deep learning approaches usually avoid the explicit feature extraction step. Classification is based on several layers learning different informative latent image representations, with different levels of granularity (i.e., a layer can learn to recognize the general round shape of the blood cell, while another layer learns to detect localized parasites inside the shape).

***Plasmodium falciparum* Detection**

To date, all studies on the detection of *Plasmodium* in blood smears have focused on the species *P. falciparum*. CNNs were applied for the first time to the identification of infected and non-infected *P. falciparum* erythrocytes by Liang et al. in 2016 [30]. Good performances were obtained (accuracy of 97.37%) after training a custom-built CNN (i.e., a CNN designed by stacking and testing successive layers) on more than 27,500 images. In comparison, the pretrained CNN variant called AlexNet [31], which typically has a more complex architecture and is known for its good performance, was less accurate for the same classification task (accuracy of 91.99%). An explanation may be that the pretraining was not specific to the task of malaria diagnosis, and the previously applied data were of a very different nature and probably less complex. Nevertheless, other studies have exploited the efficiency of pretrained models to provide more flexibility to identify parasites or to extract malarial classification features. Pretrained networks (AlexNet [31], VGG [32], and GoogLeNet [33]) were used in a study to analyze thick blood smears for *P. falciparum* ring form detection [34]. The CNNs were trained on a large number of microscopic fields, including smears from 109 patients (78 positives and 31 negatives), and scored 90% accuracy. Although this study is the only one to have analyzed thick smears on a relatively important number of patients and obtained promising results, accuracy needs to be improved for clinical applications.

Transfer learning techniques coupled with the emergence of improved pretrained models also proved useful for *P. falciparum* identification in a study by Rajaraman et al. [35]. In this study, another CNN variant model, ResNet-50 [36], offered the best performance (98.6% accuracy) compared to other pretrained models (DenseNet-121 [37] and Xception [38]). Training was performed on the malaria National Institute of Health dataset (<https://lhncbc.nlm.nih.gov/publication/pub9932>) and included nearly 27,000 equally balanced, annotated blood cells of *P. falciparum* obtained from thin blood smears. Nevertheless, when evaluating the system on various slides from individual patients, a decrease in performance was observed due to staining variations in the slides (95.9% accuracy). This remains a technical problem to be considered as blood slides may show heterogeneous staining fields easily affecting classification results. To be as close as possible to diagnostic conditions in real life, models were trained on heterogeneous images, taking into account the variability in staining results. Artificial data augmentation techniques make it possible to enrich the data available, to learn new representations, and to make a model more efficient. To improve the work performed by Rajaraman et al. [35] for *P. falciparum* identification, Rahman et al. [39] proposed an improved classification score (accuracy of 97.64%). However, data augmentation by applying several morphological transformations, such as flips, rotation, and shifting, only slightly improved accuracy (97.89%).

Generative models have the potential to create entirely new images or to modify existing images. They have become a successful and popular trend in AI. This type of model, called conditional generative adversarial networks, has been tested to generate new photorealistic images of blood cells [40]. The approach employed some recent models, such as the pix2pixHD framework [41], to work on synthetic masks of blood cell images created by the FCN-8s model [42], which were compared to real masks. A system capable of generating new representations of infected blood cells will be interesting for generating data that are often difficult to acquire due to privacy constraints

and will allow the collection of a large quantity of data for training models.

Published studies show that CNNs are a serious asset for diagnosing malaria with AI. The use of CNNs makes it possible to analyze and learn a more representative sample of the clinical routine. However, ensuring the reproducibility of the results remains a major challenge, which will require a large-scale evaluation using different imaging sources. Building collaborative databases of digitalized images for *Plasmodium* species identification is a prerequisite to addressing the variations caused by different imaging setups or differences in specimen preparation (slide realization and staining). In addition, evaluations should be performed per-sample and not per-object to be meaningful from the clinical perspective to assess identification accuracy.

Determination of *Plasmodium* Life Stages

Detecting the presence of parasites and identifying their life stages are two key steps in automated malaria diagnosis. Very few studies have addressed the determination of *Plasmodium* life stages by deep neural networks. Bashar et al. [43] evaluated a custom-built CNN for identifying *Plasmodium vivax* life stages (including gametocyte, ring, trophozoite, and schizont images) and obtained high performance scores (97.7% accuracy) with a dataset of more than 46,000 annotated images of red blood cells extracted from thin blood smears. Another study by Hung et al. [44] applied a variant object detection model, faster region-based convolutional neural network (Faster R-CNN) [45], capable of detecting and discriminating life stages of *Plasmodium vivax* from thin blood smears. An object detection model operates directly on the image, outputs bounding boxes, and classifies the detected blood cells. The model was trained on different Giemsa-stained datasets of microscopy thin blood slide images from Brazil and Thailand and was tested on a dataset from Brazil. Red blood cells were detected and separated from other blood components and sent to the AlexNet model to classify them more precisely into life stages for infected red blood cells. The system accuracy was 98%, whereas the mean accuracy obtained by two human annotators was only 72%. This evaluation

of the diagnostic expertise of humans compared to the model prediction is important to validate the performances and the potential of such a framework. However, great variability was observed between the two tested operators, suggesting a lack of expertise. This raises the problem of the gold standard and the necessity to include more operators as a reference.

A good-performing system for *Plasmodium* life-stage determination requires connecting biological expertise to data. The variability in diagnostic interpretation is an important biasing factor, and a unique reading system trained on a large volume of data, well balanced between species, will help reduce this variability. The use of CNNs makes it possible to have a more in-depth analysis of the parasites.

Mobile Applications and End-to-End Systems for Parasite Identification and Density Determination

The determination of parasite density highly depends on the performance of the developed system for parasite identification. If most studies focused on parasite identification only, a few studies addressed this important issue when evaluating automated systems for complete malaria diagnosis, from parasite identification to density determination. Coupled with AI-driven tools, the developed automated systems are of two types: mobile applications or end-to-end systems. Zhao et al. [46] published a proof-of-concept mobile application for Android smartphones combining object detection models SSD300 [47] and VGG-16. The application displayed individual images of infected blood cells and allowed authors to screen thin blood smears for *P. vivax* and *P. falciparum*. The application detected the blood cells (accuracy of 90.4%) and bound them into infected and non-infected (accuracy of 96.5%). Finally, the parasite density was computed. The authors also applied upscaling models such as fast super-resolution CNN (FSRCNN) [48] to improve the resolution and quality of images. For the training process, they used the NIH malaria dataset (<https://lhncbc.nlm.nih.gov/publication/pub9932>) [35] for *P. falciparum* detection and the Broad Institute dataset (<https://bbbc.broadinstitute.org/BBBC041>) for *P. vivax*.

detection. The *P. vivax* dataset included 1,364 blood smear images infected and uninfected, giving 80,000 individually annotated blood cells, and the rate of individual infected blood cells was 5%. Infected blood cells were automatically classified and validated by a biologist. The application also offers the possibility to load other models that can be better trained for parasite identification. Similarly, Yu et al. [49] also developed an open-source Android application for smartphone malaria diagnosis. They extended the application to the analysis of thick blood smears, a great challenge for malaria diagnosis. They evaluated the application on 200 patients (150 *P. falciparum*-infected patients and 50 non-infected patients), and performance scores ranged from 96.9% to 98.6% accuracy for thick [50] and thin [49] blood smear analysis, respectively. The application automatically quantified parasite density but also offered the possibility to help the user do it manually.

End-to-end systems for malaria diagnosis require automating both image acquisition and image analysis. This type of device generally consists of an optical system and a capture device for scanning slides, unlike smartphone applications, which are very dependent on the quality of their camera equipment but also the conditions in which the images are captured (stable position, variable light source). A system for standardizing and normalizing image acquisition will facilitate blood smear analysis and parasite identification. To address these issues, a custom-built scanner for blood thin field stack analysis was proposed by Gopakumar et al. [51]. *P. falciparum*-infected blood cells were identified by analyzing candidate regions and best focus images of objects of interest. The customized portable slide scanner showed 97.37% accuracy for parasite identification. The developed CNN was found to be more accurate, as it obtained less false positive identification of *P. falciparum* compared to a CNN without the best focus step analysis (912 FP vs. 1051 FP). However, the density estimation of the system remains far from the true estimation due to an overestimation of the presence of infected red blood cells. Commercial platforms that are available in diagnostic laboratories have emerged and are promising for malaria diagnosis. The

automated scanning microscope EasyScan Go (Motic digital[®]) analyzes thin blood slides as inputs and displays results of parasite detection and quantification for *P. vivax* and *P. falciparum* species, thanks to dedicated software, in nearly 20 min (<https://motidigitalpathology.com/easyscango/>). Another promising fully automated system, the all-in-one miLabTM platform (Noul[®]), can analyze approximately 400 to 500 thin blood smear images in 15 min (<https://noul.kr/milab-platform/>). The entire slide process is automated, including staining, and allows the bias of variability related to the preparation of the slide to be reduced. These entirely automated systems can be the next major revolution in malaria automated diagnosis.

Even if progress has been made in the field of automated microscopy malaria diagnosis, applications have not yet been implemented on a large scale in routine clinical practice. Some barriers in designing an automated malaria diagnostic system persist and have been only partially addressed. Diagnostic end-points differed from one study to another, and most of the studies have focused on the analysis of thin blood smears. End-to-end systems are promising and help to reduce the variability of data acquisition, one of the limiting factors related to the automated diagnosis of malaria. Smartphones are now available to almost everyone and allow establishing diagnostics with fast screening applications at relatively low costs. This is particularly interesting in areas with limited access to the Internet or limited access to expensive automated systems, which often requires additional training for use. Nevertheless, smartphone applications may be constrained by memory space, computational resource flexibility, or security issues.

Image-Based Automatic Classification of Mosquito Vectors

Mosquito Species Identification

Advances in AI and machine learning methods have opened the door to the automatic classification of adult mosquito vector species using recognition techniques of morphological features. Coupled with ANN, the morphological characteristics of the wings were used to classify species of

Aedes, *Culex*, and *Anopheles* mosquitoes [52]. The wing images were obtained using a digital camera coupled with a microscope at 40x magnification. All digital images were scored by a geometric morphometric analysis following 18 points called landmarks using the right wing (Fig. 3). A total of 388 specimens from 17 field-collected Culicidae species were analyzed using ANN. A correct classification was obtained in 85 to 100% of cases, depending on mosquito species. This method could receive great interest if it discriminated sibling mosquito species. However, manually annotating the images is fastidious and time consuming and is not adapted for extensive screening. In addition, wing images can be difficult to capture.

To overcome the difficulty of focusing only on the wings, other teams analyzed photographs of

whole mosquitoes using CNNs [53]. However, only 93.5% accuracy was obtained, despite a very large set of images (7,561 mosquito images) and a very low number of species to distinguish (3 species). These insufficient results may derive from a lack of robustness of input images due to the use of various cameras, angles, and resolutions. Indeed, the variability in the pose of the specimen, the scale, and lighting have been shown to be important drawbacks of image-based recognition techniques [54]. To address the drawbacks of image-based recognition techniques and to increase the robustness of input images, Park et al. [54] developed a specific method for the training of CNNs. They constructed a dataset of 3,600 images of eight mosquito species of the three major genera, *Anopheles*, *Aedes*, and *Culex*, with various

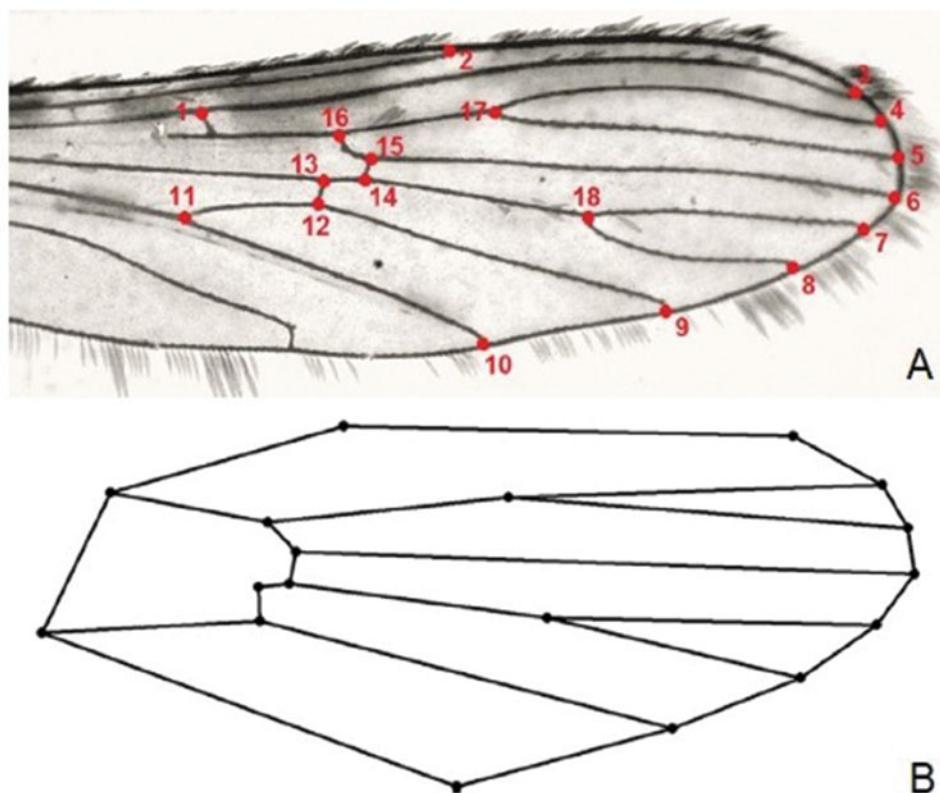


Fig. 3 Geometric morphometrics. (a) Wing of *Anopheles triannulatus* indicating the 18 landmarks used in analyses. (b) Hypothetical geometric diagram corresponding to the

wing portion examined in the study [52]. (©With courtesy of C. Lorenz)

posture, light intensity, and deformation positions (broken or missing anatomical parts) using a digital microscope. The authors achieved more than 97% accuracy by image processing using data augmentation techniques (normalizing pixel range, rotation, brightness, hue, contrast, saturation, etc.) and by fine-tuning general features using transfer learning methods (off-the-shelf features extracted from pretrained networks were reused for new target tasks).

The originality of this study was the visualization method to assess which properties of mosquitoes were used for classification, by localizing the discrimination regions of the mosquitoes that were targeted at each convolution step. Interestingly, the visual features targeted by the CNN matched the morphological key used by human experts, such as the wing venations for *Anopheles*, showing that CNN classification is supported by extracting pertinent anatomical mosquito features (Fig. 4). Nevertheless, this study highlighted the need for an important processing of mosquito images, both in terms of the number of images to acquire and in terms of preprocessing techniques, to generate a large quantity of quality data to build a performant CNN model. In addition, *Anopheles spp.* were only classified to the genus level, and

the classification of more similar species needs to be evaluated. In the future, image-based automatically classifying mosquito vectors can permit automated monitoring of mosquito species in remote field areas. Using in-field mosquito trap devices, the mosquito species can be classified automatically in real time. Such imaging systems within mosquito traps are already being developed for remote mosquito surveillance [55].

Mosquito Behavioral Patterns

An innovative AI-driven tool has recently been developed for high-throughput mosquito behavioral pattern description, the “biteOscope” [56]. It was designed to improve the comprehension of mosquito feeding behaviors around a feeding substrate (blood, sugar) and of mosquito behavioral alterations after impregnation of a surface with a chemical insect repellent. The device provides an artificial blood meal to mosquitoes by attracting them to a host mimic that they bite. For high-resolution imaging of the mosquito feeding process, the host mimic was transparent. To track each mosquito trajectory within BiteOscope, raw images were background subtracted, thresholded, and subjected to a series of morphological operations to yield binary images representing

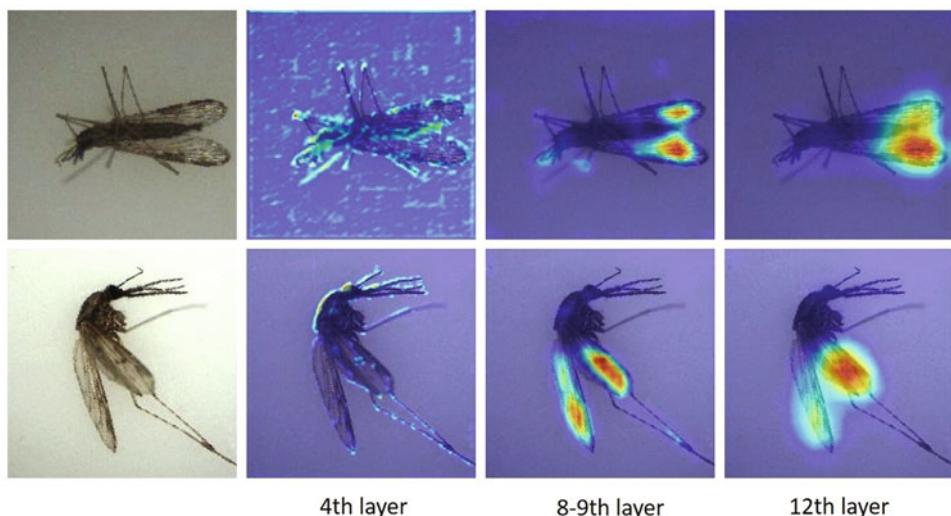


Fig. 4 Input images of two samples of *Anopheles spp.* and visualization of discriminative regions in different layers of a deep CNN: shallow (4th), middle (8th–9th) and deep

(12th) convolution layers, using a feature activation heat map [54]. (©With courtesy of J.Y. Park et al.)

mosquito bodies. Machine learning approaches were performed to extract detailed behavioral statistics from cropped images, applied as CNN input.

The locomotion, pose, biting, and feeding dynamics of *Aedes aegypti*, *Aedes albopictus*, *Anopheles stephensi*, and *Anopheles coluzzii* can be described statistically. In addition, the authors discovered how the common insect repellent DEET repels *An. coluzzii*: upon contact with their legs (Fig. 5a). This information is particularly important to design accurate mosquito vector control measures. They also observed that *An. coluzzii* approached and landed on the DEET-coated surface, but avoided prolonged contact with it. Indeed, the landing rate in the DEET area was approximately 1.9 times lower than that on the non-treated surface (Fig. 5b) and the time spent on the non-coated surface was an average of seven times longer than that on the DEET-coated surface (Fig. 5c), illustrating the precision of the statistical description.

This innovative mosquito tracking system can serve to evaluate mosquito control strategies aimed at reducing pathogen transmission by quantifying the behavioral effects of such interventions. Detailed behavioral tracking and chemical surface patterning may, respectively, enable a better knowledge of mosquito blood feeding mechanisms and ecological characteristics such as insecticide resistance. In addition, it can prove

useful for automated mosquito control and surveillance, especially in remote areas. An automated mosquito sensing system using deep learning networks has already been evaluated for the control of mosquito habitats [57]. The presence of adult mosquitoes attracted by a lure was detected by an automated system; then, mosquito detection activated the injection of a larvicide into static water to stop mosquito proliferation. This strategy targeting the larval stages can prove very useful for mosquito proliferation control in indoor urban areas such as sewage water tanks and rainfall collection facilities. These applications will constitute a real advance in the field of malaria research and mosquito control.

Characterization of *Anopheles* Biological Features Using Proteomic Tools

Only a few teams have assessed the use of MALDI-TOF MS coupled with AI tools for characterizing *Anopheles* vectors. Machine learning approaches have been tested with success to distinguish sibling *Anopheles* vector species [58]. A recent study addressed another challenge and evaluated whether MALDI-TOF mass spectra can provide suitable input for ANNs to classify the spectral patterns of *Anopheles* biological features [59]. Therefore, ANNs were coupled to

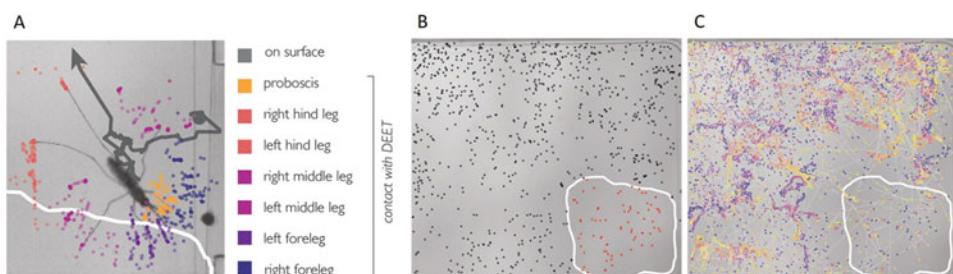


Fig. 5 DEET experiments of *An. coluzzii* repellency. (a) Body part tracking of a mosquito near the edge of the DEET-coated surface. Movement of the center of mass of the mosquito is indicated by a grey line (the start of the track is indicated by a dot, arrowhead departure). During the section of the trajectory where the mosquito is within reach of the DEET-coated area (indicated by the white line), colored dots indicate the position of the legs and

proboscis (mouth parts). (b) Landings on a substrate partly coated with 50% DEET (white line indicates DEET-coated surface). Black dots indicate landings outside the DEET area, and red dots indicate landings inside the DEET area. (c) Trajectories of mosquito movement on the surface. Dots of individual tracks are colored from purple (start of the track) to yellow (end of the track) [56]. (©With courtesy of F.J.H. Hol)

MALDI-TOF MS, and the impact on the proteome of laboratory-reared *Anopheles stephensi* mosquitoes was evaluated to predict mosquito age, blood feeding, and *Plasmodium* infection, which are *Anopheles* drivers of malaria transmission.

To build this AI application, three categories of adult *An. stephensi* mosquitoes were analyzed: (1) mosquitoes that did not receive a blood meal (unfed), (2) mosquitoes that received an uninfected blood meal on mice (fed and uninfected), and (3) mosquitoes that received a *Plasmodium berghei*-infected blood meal on mice (infected). A simplified scheme of the study design is represented in Fig. 6 for *P. berghei* infection prediction. Mosquitoes were collected at various ages and killed by freezing before processing for mass spectra acquisition. Spectra were generated from the three cohorts of mosquitoes and four body parts (head, thorax with wings, legs, abdomen). The spectra were preprocessed by smoothing and removing the baseline. Only the 100 highest peaks were fed to the ANNs to avoid background noise. Three separate ANNs (CNNs composed of

four convolutional blocks) were trained with different classification targets: age grading, past blood meal, and *P. berghei* infection.

Proteomic analysis revealed that the mass spectra were very similar, without any apparent, consistently reproducible single peak(s) correlated with each category. However, variations in peak intensity related to mosquito age, past blood meal, and *Plasmodium* infection were observed. These differences in peak intensity could create discriminant biomarkers that might be recognized by ANNs. Classification performance varied between anatomical parts, and the best prediction rates were obtained for the thorax and the legs, with 73, 89, and 78% correct predictions for age (0–10 days, 11–20 days, and 21–28 days), blood meal, and infection, respectively. This shows that the ANNs specifically recognized spectral patterns linked to *Anopheles* biology. Interestingly, the protein spectra classification was related to physiological changes, as blood meal anteriority was detected a long time after blood digestion (from 7 days to 25 days post-blood meal) from the thorax or the legs and not the

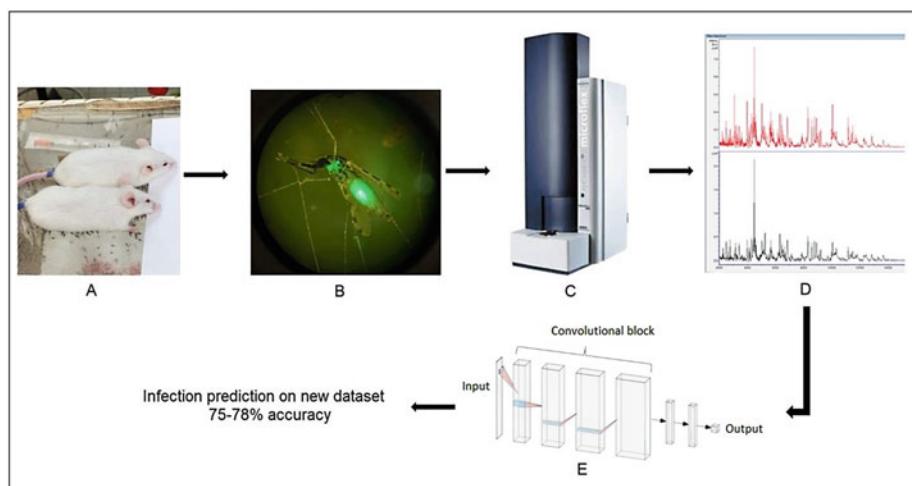


Fig. 6 Key steps to predicting *Plasmodium* infection in *Anopheles* using AI coupled to MALDI-TOF [59]. (a) Blood-feeding of *An. stephensi* mosquito on mice, either infected or not infected by *Plasmodium berghei*. (b) Control of infection by *Plasmodium berghei*. Visualization of *An. stephensi* infected with GFP-expressing *P. berghei* by fluorescence microscopy ($\times 10$ magnification), at day 24 postinfection. Fluorescence of oocysts and sporozoites

is visible in the abdomen and in the salivary glands, respectively. (c) Acquisition of MALDI-TOF MS spectra (Microflex LT Bruker[®]). (d) Categories of mass spectra protein profiles to be discriminated. Red color indicates a spectrum of a mosquito fed and infected. Black color indicates a spectrum of a mosquito fed and uninfected. (e) Machine learning using a convolutional neural network

abdomen. Similarly, the neural network could distinguish between infected and uninfected *Anopheles* from the legs, which are not supposed to host the parasite for a long time, suggesting detecting protein changes following *Plasmodium* infection such as the immune response rather than detecting the parasite.

This proof of concept illustrates that MALDI-TOF MS coupled to ANNs can provide new tools for vector control by estimating the proportions of *Plasmodium*-infected, blood-fed, and old mosquitoes. These proportions are expected to decrease under effective vector control measures. The age-grading approach is particularly interesting, as age cannot be provided by DNA analysis, and the reduction in mosquito age populations is one of the most effective measures for reducing malaria transmission [19]. Future studies must assess the field validity of this new approach in wild-caught adult *Anopheles*.

Conclusion

Researchers from computer science, life science, and medical science have addressed the challenges of malaria diagnosis and control using AI during the last decade. In the field of malaria diagnosis, many promising AI applications are under development. In the future, they can bring efficient and accurate diagnostic tools, having a real impact on clinical management, by revolutionizing the urgent diagnosis of malaria. Nevertheless, a standardized universal system for automated malaria diagnosis will have to consider the variability resulting from staining methods and the quality of smear preparation for microscopic blood smear images. A fair comparison of the performances between different algorithms was not always possible due to the use of different evaluation metrics and due to the lack of reference benchmark data. In addition, most systems have been developed and tested using a relatively small number of patients due to difficulties in acquiring and annotating a large number of images. However, computer-assisted diagnosis and mobile applications provide a significant opportunity for data collection and for promoting the creation of centralized collaborative red blood cell databases.

The generated data will permit researchers to improve model training and enrich the representations and cell variability of malaria, especially when these data are scarce. In the field of mosquito surveillance and control, few researchers have taken up the challenge of developing exciting AI applications. They can not only identify mosquito species but also characterize mosquito biology. If field-validated, these promising approaches will facilitate the epidemiological monitoring of malaria vectors and the saving of resources by preventing or reducing malaria transmission.

Cross-References

► AIM and mHealth, Smartphones and Apps

References

1. Manguin S, Carnevale P, Mouchet J, Coosemans M, Julvez J, Richard-Lenoble D, et al. Biodiversity of malaria in the world. Manguin S, Carnevale P, Mouchet J, editors. Paris: John Libbey Eurotext; 2008. 464–478 p.
2. White NJ, Pukrittayakamee S, Hien TT, Faiz MA, Mokuolu OA, Dondorp AM. Malaria. Lancet. 2014;383(9918):723–35.
3. World Malaria Report. 20 years of global progress and challenges. Geneva: World Health Organization; 2020. p. 2020.
4. Global Vector Control Response 2017–2030. Geneva: World Health Organization; 2017. Licence: CC BY-NC-SA 3.0 IGO.
5. Michalakis Y, Renaud F. Malaria: evolution in vector control. Nature. 2009;462(7271):298–300.
6. Maze MJ, Bassat Q, Feasey NA, Mandomando I, Musicha P, Crump JA. The epidemiology of febrile illness in sub-Saharan Africa: implications for diagnosis and management. Clin Microbiol Infect. 2018;24(8):808–14.
7. Wongsrichanalai C, Barcus MJ, Muth S, Sutamihardja A, Wernsdorfer WH. A review of malaria diagnostic tools: microscopy and rapid diagnostic test (RDT). Am J Trop Med Hyg. 2007;77(Suppl 6):119–27.
8. Mathison BA, Pritt BS. Update on malaria diagnostics and test utilization. J Clin Microbiol. 2017;55(7):2009–17.
9. Mayengue PI, Kouhounina Batsimba D, Dossou-Yovo LR, Niama RF, Macosso L, Pembet Singana B, et al. Evaluation of routine microscopy performance for malaria diagnosis at three different health centers in Brazzaville, Republic of Congo. Malar Res Treat. 2018;2018:4914358.

10. Gwer S, Newton CRJC, Berkley JA. Over-diagnosis and co-morbidity of severe malaria in African children: a guide for clinicians. *Am J Trop Med Hyg.* 2007;77(Suppl 6):6–13.
11. A-Elgayoum SME, El-Feki AEKA, Mahgoub BA, El-Rayah EA, Giha HA. Malaria overdiagnosis and burden of malaria misdiagnosis in the suburbs of central Sudan: special emphasis on artemisinin-based combination therapy era. *Diagn Microbiol Infect Dis.* 2009;64(1):20–6.
12. Mangun S, Garros C, Dusfour I, Harbach RE, Coosemans M. Bionomics, taxonomy, and distribution of the major malaria vector taxa of *Anopheles* subgenus *Cellia* in Southeast Asia: an updated review. *Infect Genet Evol.* 2008;8(4):489–503.
13. Wolk DM, Clark AE. Matrix-assisted laser desorption time of flight mass spectrometry. *Clin Lab Med.* 2018;38(3):471–86.
14. Normand A-C, Becker P, Gabriel F, Cassagne C, Accoceberry I, Gari-Toussaint M, et al. Validation of a new web application for identification of fungi by use of matrix-assisted laser desorption ionization-time of flight mass spectrometry. *J Clin Microbiol.* 2017;55(9):2661–70.
15. Murugaiyan J, Roesler U. MALDI-TOF MS profiling—advances in species identification of pests, parasites, and vectors. *Front Cell Infect Microbiol.* 2017;7:184.
16. Yssouf A, Almeras L, Raoult D, Parola P. Emerging tools for identification of arthropod vectors. *Future Microbiol.* 2016;11(4):549–66.
17. Nabet C, Kone AK, Dia AK, Sylla M, Gautier M, Yattara M, et al. New assessment of *Anopheles* vector species identification using MALDI-TOF MS. *Malar J.* 2021;20(1):1–17.
18. Lachaud L, Fernández-Arévalo A, Norman AC, Lami P, Nabet C, Donnadieu JL, et al. Identification of *Leishmania* by matrix-assisted laser desorption ionization-time of flight (MALDI-TOF) mass spectrometry using a free web-based application and a dedicated mass-spectral library. *J Clin Microbiol.* 2017;55(10):2924–33.
19. Johnson BJ, Hugo LE, Churcher TS, Ong OTW, Devine GJ. Mosquito age grading and vector-control programmes. *Trends Parasitol.* 2020;36(1):39–51.
20. Pollak JJ, Houri-Yafin A, Salpeter SJ. Computer vision malaria diagnostic systems – progress and prospects. *Front Public Health.* 2017;5(August):1–5.
21. Rehman A, Abbas N, Saba T, Mehmood Z, Mahmood T, Ahmed KT. Microscopic malaria parasitemia diagnosis and grading on benchmark datasets. *Microsc Res Tech.* 2018;81(9):1042–58.
22. Schwalbe N, Wahl B. Artificial intelligence and the future of global health. *Lancet.* 2020;395(10236):1579–86.
23. Savkare SS, Narote SP. Automated system for malaria parasite identification. In: 2015 International conference on communication, information and computing technology (ICCICT). Mumbai; 2015.
24. Abdul Nasir AS, Mashor MY, Mohamed Z. Segmentation based approach for detection of malaria parasites using moving k-means clustering. In: 2012 IEEE-EMBS conference on biomedical engineering and sciences. Langkawi; 2012. p. 653–8.
25. Khan NA, Pervaiz H, Latif A, Musharaff A. Unsupervised identification of malaria parasites using computer vision. In: 2014 11th international joint conference on computer science and software engineering (JCSSE). Chon Buri; 2014. p. 263–7.
26. Hearst MA, Scholkopf B, Dumais S, Osuna E, Platt J. Trends and controversies - Support vector machines, 13 (4). In: IEEE Intelligent systems and their applications; 1998. p. 18–28.
27. Das DK, Maiti AK, Chakraborty C. Automated system for characterization and classification of malaria-infected stages using light microscopic images of thin blood smears. *J Microsc.* 2015;257(3):238–52.
28. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med.* 2018;24(9):1342–50.
29. Zhang Z, Chen P, McGough M, Xing F, Wang C, Bui M, et al. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nat Mach Intell.* 2019;1(5):236–45.
30. Liang Z, Powell A, Ersoy I, Poostchi M, Silamut K, Palaniappan K, et al. CNN-based image analysis for malaria diagnosis. In: 2016 IEEE international conference on bioinformatics and biomedicine (BIBM). Shenzhen; 2016. p. 493–6.
31. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM.* 2017;60(6):84–90.
32. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: The 3rd International Conference on Learning Representations (ICLR 2015); arXiv:14091556v6.
33. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR). Boston; 2015. p. 1–9.
34. Mehanian C, Jaiswal M, Delahunt C, Thompson C, Horning M, Hu L, et al. Computer-automated malaria diagnosis and quantitation using convolutional neural networks. In: 2017 IEEE international conference on computer vision workshops (ICCVW). Venice; 2017. p. 116–25.
35. Rajaraman S, Antani SK, Poostchi M, Silamut K, Hossain MA, Maude RJ, et al. Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. *PeerJ.* 2018;6:e4568.
36. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). Las Vegas; 2016. p. 770–8.
37. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR). Honolulu; 2017. p. 2261–9.
38. Chollet F. Xception: deep learning with depthwise separable convolutions. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR). Honolulu; 2017. p. 1800–7.

39. Rahman A, Zunair H, Rahman MS, Yuki JQ, Biswas S, Alam MA, et al. Improving malaria parasite detection from red blood cell using deep convolutional neural networks. arXiv. 2019; 190710418.
40. Bailo O, Ham D, Shin YM. Red blood cell image generation for data augmentation using Conditional Generative Adversarial Networks. In: 2019 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW); Long Beach; 2019. p. 1039–48.
41. Wang TC, Liu MY, Zhu JY, Tao A, Kautz J, Catanzaro B. High-resolution image synthesis and semantic manipulation with conditional GANs. In: 2018 IEEE/CVF conference on computer vision and pattern recognition. Salt Lake City; 2018. p. 8798–807.
42. Long J, Shelhamer E, Trevor D. Fully convolutional networks for semantic segmentation. In: 2017 IEEE transactions on pattern analysis and machine intelligence; 2017. p. 640–51. <https://doi.org/10.1109/TPAMI.2016.2572683>
43. Bashar MK. Automated classification of malaria parasite stages using convolutional neural network-classification of life-cycle stages of malaria parasites. In: 2019 Proceedings of the 3rd international conference on vision, image and signal processing; Vancouver, 2019. p. 1–5.
44. Hung J, Carpenter A. Applying faster R-CNN for object detection on malaria images. In: 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW); arXiv:1804.09548v2; 2017. p. 56–61.
45. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. In: 2015 IEEE transactions on pattern analysis and machine intelligence; arXiv:1506.01497v3; 2015. p. 1–9.
46. Zhao OS, Kolluri N, Anand A, Chu N, Bhavaraju R, Ojha A, et al. Convolutional neural networks to automate the screening of malaria in low-resource countries. PeerJ. 2020;8:e9674.
47. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, et al. SSD: single shot multibox detector. In: European conference on computer vision (ECCV); arXiv:1512.02325v5; 2016. p. 21–37.
48. Dong C, Loy CC, Tang X. Accelerating the super-resolution convolutional neural network. In: 2016 Proceedings of European conference on computer vision (ECCV); arXiv:1608.00367v1; 2016. p. 391–407.
49. Yu H, Yang F, Rajaraman S, Ersoy I, Moallem G, Poostchi M, et al. Malaria Screener: a smartphone application for automated malaria screening. BMC Infect Dis. 2020;20(825):1–8.
50. Yang F, Yu H, Silamut K, Maude RJ, Jaeger S, Antani S. Smartphone-supported malaria diagnosis based on deep learning. In: Suk HI, Liu M, Yan P, Lian C, editors. Machine learning in medical imaging. MLMI 2019. Lecture Notes in Computer Science, vol. 11861. Springer; 2019.
51. Gopakumar GP, Swetha M, Sai Siva G, Sai Subrahmanyam GRK. Convolutional neural network-based malaria diagnosis from focus stack of blood smear images acquired using custom-built slide scanner. J Biophotonics. 2018;11(3):1–17.
52. Lorenz C, Ferrando AS, Suesdek L. Artificial Neural Network applied as a methodology of mosquito species identification. Acta Trop. 2015;152:165–9.
53. Motta D, Bandeira Santos AA, Souza Machado BA, Vicente Ribeiro-Filho OG, Arriaga Camargo LO, Valdenegro-Toro MA, et al. Optimization of convolutional neural network hyperparameters for automatic classification of adult mosquitoes. PLoS One. 2020;15(7):1–30.
54. Park J, Kim DI, Choi B, Kang W, Kwon HW. Classification and morphological analysis of vector mosquitoes using deep convolutional neural networks. Sci Rep. 2020;10(1):1–12.
55. Goodwin A, Glancey M, Ford T, Scavo L, Brey J, Heier C, et al. Development of a low-cost imaging system for remote mosquito surveillance. Biomed Opt Express. 2020;11(5):2560–9.
56. Hol FJH, Lambrechts L, Prakash M. BiteScope: an open platform to study mosquito blood-feeding behavior. elife. 2020;9:e56829.
57. Kim K, Hyun J, Kim H, Lim H, Myung H. A deep learning-based automatic mosquito sensing and control system for urban mosquito habitats. Sensors (Basel). 2019;19(12):2785.
58. Müller P, Pflüger V, Wittwer M, Ziegler D, Chandre F, Simard F, et al. Identification of cryptic *Anopheles* mosquito species by molecular protein profiling. PLoS One. 2013;8(2):e57486.
59. Nabet C, Chaline A, Franetich JF, Brossas JY, Shahmirian N, Silvie O, et al. Prediction of malaria transmission drivers in *Anopheles* mosquitoes using artificial intelligence coupled to MALDI-TOF mass spectrometry. Sci Rep. 2020;10(1):11379.



Artificial Intelligence in Infection Biology

98

Artur Yakimovich

Contents

Introduction	1370
Aim in the Infection Biology	1370
Computer Vision in Infection Biology on Nano- and Microscale	1371
Computer Vision in Infection Biology on Mesoscale and Aspects of Temporal Dimensions	1373
Computer Vision in Infection Biology on Macroscale and Digital Biomarkers	1373
Artificial Intelligence in Molecular Infection Biology	1374
Conclusions	1375
References	1376

Abstract

Recent methodology advances in infection biology suggest a great potential of artificial intelligence. As the field paces into the digital age, data-driven techniques for analysis of pathogens are being widely embraced. From computer vision in microscopy to sequence analysis in genomics of pathogens, these advances allow to improve speed, specificity, and sensitivity of infectious disease diagnostics. Advances in host-pathogen interaction powered by artificial intelligence promise novel therapies. In this chapter, the background of infection biology field and types of data it deals with are introduced. Across the spatial and temporal scales, specific examples of how

the classic techniques are getting an artificial intelligence rehaul are provided. Furthermore, novel opportunities that artificial intelligence, machine learning, and deep learning are offering for the field are discussed. These include label-free infection detection, causal biological mechanism analysis, and AI-assisted design and identification of novel antivirals and antibiotics. Finally, a perspective on limitations and opportunities of artificial intelligence techniques in infection biology is discussed.

Keywords

Infection biology · Virus · Bacteria · Pathogens · Artificial intelligence · Infectious diseases · Machine learning · Deep learning · Microscopy · Biomedical imaging

A. Yakimovich (✉)

Artificial Intelligence for Life Sciences CIC, London, UK

e-mail: ayakimovich@ails.institute

© Springer Nature Switzerland AG 2022

N. Lidströmer, H. Ashrafiyan (eds.), *Artificial Intelligence in Medicine*,
https://doi.org/10.1007/978-3-030-64573-1_105

1369

Introduction

Infectious diseases have made an immense socio-economic impact throughout the history of humanity. Traditions and cultures were shaped by diseases through effects in human populations and livestock. Advances in infection biology left noticeable hallmarks in science and medicine including works of Jenner, Lister, Koch, Pasteur, and Fleming [1, 2]. It would hardly be an overstatement to say that these achievements define the very fabric of our modern society. Yet, the ability to fight the microscopic parasites wasn't always a given and, in the face of the newly emerging infections, may not always continue to be [3].

Detecting and preventing the ever-changing microbes and viruses requires constant survey and research. However, today these efforts are greatly facilitated by the achievements of the past. Standing on the shoulders of giants, we are able to borrow from the basic concepts of biology, biochemistry, biophysics, genetics, and bioinformatics to devise new ways of fighting infectious disease. Discovered through centuries of empiric observations, remarkably, these concepts often hold today. Infection often meant deviation from the norm; hence the effect of pathogens on their hosts could be observed – i.e., learning from the data. This way of thinking about infection biology bears parallels to the machine learning (ML) and representation learning techniques. In this chapter, we will explore how artificial intelligence (AI) explores these parallels to give science and medicine an edge in this perpetual armament race.

Processes studied by infection biology exist on various scales and involve various agents. To capture this, complexity on a fairly abstract level lets us describe these agents across the scales, as well as the processes they are involved in (Fig. 1). In the first dimension of the scales, these agents exist on a spatial scale from molecules to organisms (Fig. 1a). In the second, these agents change in time on a broad range of temporal scales from milliseconds to days and, hence, have a temporal dimension. Throughout these dimensions, these multiscale agents engage in a plethora of interdependent processes (Fig. 1b), which, in

turn, may have outcome on each of the scales of the spatiotemporal range. In other words, infection biology studies entities from small to large and from fast to slow, and minute differences may dramatically change a life-or-death outcome for an organism.

Noteworthy, there are two distinct foci of infection biology. First is the focus on the pathogen as a biological entity itself. The second is the host-pathogen interaction [4] focus – i.e., on the interactions between the pathogen and the host, involving multiple stages of the pathogen's lifecycle (see Fig. 1b). Pathogen focus may, for example, relate to the analysis of its shape or genome. Host-pathogen interaction focus relates to interactions between genes and molecules of the host cell and the pathogen molecules. Such interactions may include mediating various host mechanisms exploitation by the pathogen. A great example of such exploitations is pathogen entry into the host cell [5].

Here, we will review the role microscopy and computer vision can play for AI in infection biology, akin to the role microscopy played first explorers of the microbiological world [6]. Next, we will discuss strengths and limitations of using various ML and deep learning (DL) techniques [7], which gained popularity in recent years. Specifically, we will focus on their application to pathogen detection or analysis of host-pathogen interactions. Finally, we will explore AI application to molecules in the context of infection biology. This chapter is not going to address epidemiology aspects of infections, which are reviewed in [8]. Furthermore, we will omit many important, yet complex, aspects of immunology.

Aim in the Infection Biology

Pathogens span a broad range of sizes ranging from nanometers to micrometers. Curiously, this range of scales is also divided by the so-called diffraction limit (see Fig. 1). This limit splits pathogens into resolvable by optical imaging and resolvable largely by electron microscopy [9]. When it comes to viruses, for example, the impact of electron microscopy on the field of

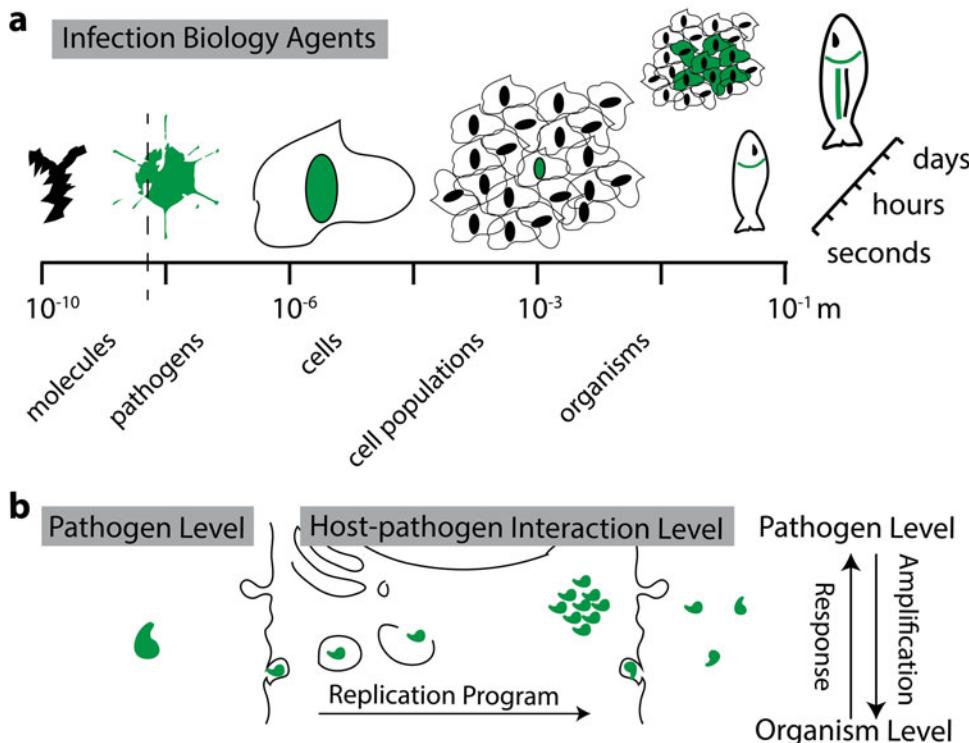


Fig. 1 Abstract description of agent space and processes studied by infection biology. (a) Agents involved and their respective positions on the spatiotemporal scale. Here green color indicates infected state or infectious agent. Dashed line denotes approximate location of the diffraction limit on the spatial scale. (b) Abstract depiction

of the processes the agents are involved in. Here curved lines are used to schematically depict host cell, endocytic cell entry, replication, and egress of the pathogen progeny. A generic pathogen is depicted in green. Any immune interactions, here, are depicted as “Response”

virology is difficult to overestimate. Ability to resolve virus particles allowed researchers to “put a face” on the virus entity. Yet, even larger pathogens like pathogenic bacteria and parasites clearly require special equipment to be visualized.

Conversely, the effect pathogens exert on their host is best appreciated on a larger scale. To be able to encompass the broad application of AI for the topic of infection biology, in this chapter, we will begin our journey with vision on nano- and microscales, proceed to the mesoscale, and wrap up with the macroscale. In mesoscale, we will mention some aspects of temporal dimensions of infection biology, although, wholeheartedly, this aspect may well deserve its own chapter. Finally, making a full circle, we will return to molecules and the role AI can play for the molecular techniques in infection biology.

Computer Vision in Infection Biology on Nano- and Microscale

In the past decade, the field of computer vision has been revolutionized by the convolutional neural networks (CNNs) and similar artificial neural network methods enabled by the gradient-based learning [10]. However, these are most powerful in situations where datasets are very large, while data annotation is trivial and unambiguous. It is easy to imagine how this can be problematic for rare pathogen analysis at a nano- and microscale. Consider, for example, annotation of handwritten digits – where ground truth is obvious, versus recognizing obscure virus particles in the gray scale electron micrograph. Such annotation task for a human specialist often requires years of training. Furthermore, consensus among

specialists often leaves much to be desired. In such cases, a simple rule-based analysis approach is often preferable [11].

Additionally, the nature of biomedical and other scientific images is often quite different compared to conventional image computer vision is dealing with. Scientific images are typically acquired in very standardized conditions using highly accurate equipment developed by a handful of vendors. Magnifications or the fields of view are often fixed. The very design of the scientific images is aimed at leaving very little to a mere chance: depth of field is shallow, shape and size distortion is minimized, background is standardized, and the perspective is eliminated. In case of so-called functional imaging (opposed to label-free imaging), entities of interest may be stained through chemical dyes and contrast compounds providing an experimental solution to problems like object segmentation. Hence, rule-based algorithms have dominated the field for a long time.

One remarkable breakthrough in the field, with respect to ANN, came from the U-net architecture addressing the segmentation problem in the label-free imaging modalities [12]. This architecture is exploiting highly standardized scales in the scientific image to draw various representations of such fields of view. In turn, pathogens in general and viruses in particular often have a very distinct morphology. This peculiarity may be exploited both for novel imaging methods (using pathogens as fiducials) and for better understanding structure and function relationships with the pathogens [13]. Unsurprisingly U-net architecture was shown to be of a great use for virus particle recognition in electron microscopy images [14].

Beyond segmentation, computer vision and AI opened another avenue in the label-free imaging, namely, performing quantitation of biological entities based on their distinctive morphology rather than specific labeling. Examples include gram stain classification from bacterial morphology in blood culture [15] and use of holographic imaging modalities accompanied by CNNs for screening of anthrax spores [16], as well as Raman spectroscopy-based bacteria pathogenicity interpretation [17]. Noteworthy, methodologies utilized

in these works all use CNNs, yet they differ significantly in the specific implementation. Jo et al. employ a relatively shallow CNN architecture they term HoloConvNet consisting of merely three consecutive CNN layers. At the same time Smith and colleagues use a deep residual network architecture based on Inception v3 equipped with 23,885,392 trainable parameters. Generally speaking, the expressive capacity of ANN grows with the number of trainable parameters. However, the annotated data requirements grow with the increase of expressive capacity. High-capacity networks tend to overfit on smaller datasets. Perhaps it was the data availability that dictated the choice of ANN architecture for these papers. Another aspect of residual architecture, however, is their ability to tune the expressive capacity through so-called skip-connections between the layers. That said, deeper architectures typically require more memory on GPU devices for training, hence likely a more expensive device at hand – an important fact to consider while designing a project.

Another opportunity for AI in the infection biology on a nano- and microscopic scales comes from analysis of the host-pathogen interactions. In these applications, algorithms like CNNs can assist with understanding how host cell molecules interact with the intruding pathogens. If the detection of interacting and non-interacting pathogens is formulated as a classification problem, the implicit unsupervised component of CNNs may be employed to learn and visualize previously unknown characteristic properties of the image [18, 19]. Furthermore, certain aspects of data requirement of ANN architectures may be relaxed by employing so-called transfer learning approaches, which work surprisingly well in a cross-domain application [20]. Taking this notion further, CNNs may be readily employed to decipher unknown host-pathogen mechanisms, like in the example of infected cell lysis by some human adenoviruses [21]. For example, Andriasyan et al. employ residual CNN accompanied by saliency techniques (so-called class activation maps (CAMs)) to identify elusive phenotype of cell infected by human adenovirus yet unable to spread the virus. Using these techniques accompanied by laser-ablative infected cell probing,

authors unveil that the build-up of intranuclear pressure plays a key role in human adenovirus cell lysis.

Computer Vision in Infection Biology on Mesoscale and Aspects of Temporal Dimensions

Effects of pathogens are rarely limited to the nano- and microscale. On the path of infection spread, pathogens transcend the limits of individual cells. In monolayers of cultured cells, viruses form distinct clonal lesions called viral plaques. Computer vision algorithms can greatly assist in quantitation of these infection phenotypes [22]. On a slightly larger scale, individual bacterial cells can be grown in colonies on dishes covered with solidified growth medium. This technique is widely used throughout microbiology and molecular biology (e.g., for molecular cloning of DNA), as well as in clinical diagnostics. Quantitation of such colonies, however, is often done manually in a slow and cumbersome process. Recently, CNN-based methods for automated colony counting [23] and classification [24] in digital microbiology imaging have been proposed. Noteworthy, however, Ferrari and colleagues took great care in standardizing the image acquisition condition for their agar plates. This suggests that their model may not be immediately applicable for everyone and may require retraining. Nevertheless, these works clearly demonstrate viability of CNN application in infection biology at a mesoscale and suggest an enormous potential for improving speed and precision of infectious disease diagnostics.

Effects of pathogens in the populations of cells do not occur in an instance. Depending on the duration of the pathogens' lifecycle, such effects may be taking place on a temporal scale ranging between hours and days to even weeks. Curiously, quantitating the rate of change in such processes can bear a great deal of information about infection spread. For example, tracking migrations of vaccinia virus infected cells has been recently used to unveil that the pathogen hijacks host cell EGFR signaling mechanism to enhance virus

spread through cell motility [25]. A curious approach taken in this study emphasizes how important mechanistic cues may be retrieved from swarm-like infected cell behavior at a mesoscale.

Finally, employing computer vision methodology to measure infection at the mesoscale paves the way to new approach to scientific instruments. For example, recent advances in mobile photography have seen a tremendous increase in detail and quality (i.e., pixel resolution in cameras and postprocessing). With minor modifications, a smartphone can literally be converted into a low-cost microscope. In turn AI algorithms can facilitate analysis, like bacteria-type detection [26]. Furthermore, the farther end of the mesoscale allows phenotypes to be large enough to be visible with a naked eye – a scale where mobile photography can provide highly available and reproducible way of digitalizing research results.

Computer Vision in Infection Biology on Macroscale and Digital Biomarkers

When infections expand beyond the mesoscale of cellular neighborhood, they spread to tissues – or the macroscale, as termed here. The adoption of imaging in clinics has truly transformed medicine in recent years: digitalization of X-ray, computed tomography (CT), and magnetic resonance imaging (MRI) are more often becoming common tools to extend the diagnostic picture. Detecting tissue abnormality can be a powerful tool in medicine. One remarkable example is the usage of retinal scans to detect macular diseases. Furthermore, retinal scan data may provide cues to a whole range of systemic pathologies. Unsurprisingly, CNNs and DL in general are spearheading analytics in this aspect [27]. Beyond analytics formulated as pathology classification problem, great advances have been achieved in using AI for signal processing. Training model on pre- or low-dosed contrast-enhanced brain MRI combined with the full-dose images as ground truth have been shown to allow for reducing the amount of contrast compound required for functional brain MRI imaging [28]. This, in turn, may

allow for safer diagnostics through decreasing adverse events' frequency connected to contrast compound.

Similarly, CT and MRI imaging are advancing diagnostics of infectious diseases. Although detecting individual pathogens in whole organ scans would be impossible due to the massive difference of scales, one can readily observe the damage they cause. These techniques bridge the medical gap between the infection and disease. While infection is the mere fact of presence and replication of the pathogen in cells and tissues, disease is the condition caused by the infection itself or response to the infection. Bridging this gap is important to broaden our understanding of pathogens and infectious diseases in general, as not all infections lead to the development of the disease. One recent and illustrative example is the SARS-CoV2 virus infection being the causative agent of coronavirus infectious disease 19 (COVID-19). While COVID-19 is characterized by the manifestation of a severe acute pneumonia, as of this writing, the general consensus stands that by far not every infection of SARS-CoV2 leads to the development of the COVID-19 condition. Hence, molecular diagnostics of SARS-CoV2 infections like polymerase chain reaction (PCR) remains insufficient for the obtaining a clinical picture for COVID-19 patients. However, this gap can be filled by the clinical picture obtained from imaging and accompanied by DL [29].

Beyond imaging, other AI techniques may be of great use in clinics for infectious disease diagnostics. For example, reinforcement learning (RL) technique holds an immense potential for optimizing treatment strategies by taking into account a myriad of input features measured by the various medical sensors available today. Signals with high temporal resolution from wearable devices and advanced medical equipment used in the intensive care unit wards are a valuable source of so-called digital biomarkers. While gigabytes of data from such sensors are recorded, they rarely play a big role in today's diagnostics, which is wired for more conventional bulk readouts. Furthermore, such big diagnostic data by itself may lead to information overload for clinicians and hamper fast clinical decisions. However, if such rich

clinical data source is accompanied by an advance analytics system, it may prove remarkably useful. One example of such synergy is the use of RL to select optimal treatment strategies for sepsis in intensive care [30]. In this study, authors used a set of 48 variables, including laboratory values, demographics, vital signs, comorbidity status, fluids, and vasopressors received to power their AI system. These data were recorded as multi-dimensional discrete time series with 4-h time steps and with a total of up to 72 h of measurements for each patient. All this allowed to estimated time of onset of sepsis at the level of human physician. Furthermore, the system developed by Komorowski et al. occasionally outperformed optimal vasopressor dosage suggestion compared to the dosage proposed by human doctors, suggesting potential improvement of the state of the art.

Yet, simply matching the performance of human physician in diagnosis and treatment of infectious diseases is not the only promising advancement of AI. The digital biomarker paradigm allows to bring novel information sources to be used for diagnostic purposes. For example, Laguarta and colleagues have recently proposed a CNN-powered strategy to detect COVID-19 from a forced-cough audio recording obtained through a cell phone [31]. Using data of over 5000 COVID-19 patients and ResNet-based architecture, authors developed a model capable of detecting COVID-19 from audio with a sensitivity of 98.5% with a specificity of 94.2%. This and other examples of digital biomarker application in infectious disease diagnostics bear a great inspiration of utilizing previously untapped sources for diagnostics.

Artificial Intelligence in Molecular Infection Biology

While computer vision may be one of the most obvious applications of AI methodology in infection biology, it would be a gross omission to ignore the great advances AI is driving in pathogens genetics and molecular infection biology. For example, DL may be readily applied as pattern recognition approach to identify genetic

variations responsible for higher pathogenicity in either host or pathogen genomes [32]. Beyond genetic variations, DL can assist with DNA methylation analysis, base calling and SNP analysis [33], transcription analysis, RNA analysis, as well as DNA accessibility analysis (reviewed in [32]). However, opportunities for AI in infection biology are not limited to improving conventional genomics of individual host and a pathogen. Whole microbiome may be taken into account in the clinical setting through the approach known as clinical metagenomics [34].

More often than not, molecular infection biology deals with analysis of patterns in sequences. Hence, hand in hand with CNNs, ANN approaches like recurrent neural networks (RNN), for example, architecture known as long short-term memory (LSTM) RNN [35], are widely used in the field [32]. In fact, in some cases, these architectures can also be synergistically combined. For example, the work of Veltre et al. uses consecutive combination of CNN and LSTM layer in a single architecture aimed at antimicrobial peptide recognition [36]. Authors show superiority of their approach to either layers applied individually.

Our overview would be incomplete without mentioning the incredible innovation DL has demonstrated for the chemical synthesis [37]. This, in turn, allows to appropriate the AI methodology for drug discovery. Specifically, Stokes and colleagues have recently proposed a DL approach to antibiotic discovery [38]. In their work, authors use a graph neural network and data from Drug Repurposing Hub, as well as ZINC15 databases to predict a novel antibiotic. Remarkably, Stokes et al. identify a small molecule Halicin (5-[(5-nitro-1,3-thiazol-2-yl)sulfanyl]-1,3,4-thiadiazol-2-amine) as a potent antibiotic candidate. Furthermore, they demonstrate that the molecule readily shows activity in mice – a remarkable step for an *in silico* methodology. Similar methodologies also show promise in the scarce antiviral space. One remarkable advantage the use of AI may provide is the rapid reaction speed to the emerging infection. For example, using a DL model of drug-target interaction, Beck et al. identified potential antivirals for the pandemic SARS-CoV2 from a broad range of

commercially available antivirals. In this study, the team has developed molecule transformer-drug target interaction architecture capable of utilizing molecular information represented as simplified molecular input-line entry system (SMILES) code. Authors have identified human immunodeficiency virus drug as a putative candidate against SARS-CoV2 [39].

Yet, one of the biggest criticisms of large DL models remains their poor interpretability. For domains like drug discovery, such limitations may be critical for the viability of the approach. To this end, Yang and colleagues have recently proposed a white-box ANN approach [40]. To achieve this, Yang et al. counter-screened a number of metabolites against bactericidal antibiotics in *Escherichia coli*. Next, authors simulated the corresponding states of the metabolites using a genome-scale metabolic network model. This allowed the team not only to identify novel molecules but also obtain causal mechanisms underlying drug efficacy.

Conclusions

In this AIM chapter, we have discussed the multi-scale nature of infection biology as well as multi-faceted approaches required to detect, study, and ultimately treat infectious diseases. We have dived into the origins of the field of infection biology. Also, we have introduced the space of agents involved in the infection biology as well as pathogen-centric and host-pathogen-interaction-centric views in the field. We have reviewed recent advances in domain specific AI pushing the boundaries of previously available infection biology methodologies. While these advances are truly breath-taking, the “No Free Lunch” theorem dictates that all these advances come at a cost and must be perceived with a grain of salt.

Indeed, a lot of AI approaches in the field are not immediately applicable to any conceivable problem at hand. These large models require immense annotated datasets to train, as well as large computational resources. A typical size of a dataset used in DL may count tens or hundreds or thousands of individual data points (e.g., images), as well as days of training on expensive

GPU-enabled computers. Furthermore, for these datasets to be suitable for supervised ML, they need to be annotated by domain experts of each individual narrow field, making the annotated datasets scarce. Further complicating the adoption of the AI methodologies in the infection biology, data are often imbalanced – i.e., only few positive examples are often overwhelmed by the vastness of negative examples. The abovementioned challenges with DL models' interpretability additionally hamper the motivation of applying AI to these problems. Needless to say, all these present great challenges for the field to overcome. However, some of these challenges are not unique to AI in infection biology.

Indeed, recent advances in semi- and self-supervised ML/DL as well as work on universal representations promise to address the scarcity of annotations. Advances in techniques of data augmentation and synthetic data give hope to tackle dataset imbalance. Use of transfer learning [19] and broader availability of GPU resources could hopefully address the computational costs. Finally, the white-box causal AI model design [40] as well as approaches to CNN saliency [18, 19] could shed light on the inner workings of the advanced models – a necessary condition for their wide adoption in the domain. All these opportunities provide a promise of the wide adoption of AI in infection biology driving the field into the new age.

References

- Buchholz K, Collins J. The roots – a short history of industrial microbiology and biotechnology. *Appl Microbiol Biotechnol*. 2013;97(9):3747–62.
- Riedel S. Edward Jenner and the history of smallpox and vaccination. In: Baylor University Medical Center proceedings, vol. 1. Taylor & Francis; 2005. p. 21–5.
- McMichael AJ. Environmental and social influences on emerging infectious diseases: past, present and future. *Philos Trans R Soc Lond B Biol Sci*. 2004;359(1447):1049–58.
- Casadevall A, Pirofski L-A. Host-pathogen interactions: basic concepts of microbial commensalism, colonization, infection, and disease. *Infect Immun*. 2000;68(12):6511–8.
- Yamauchi Y, Helenius A. Virus entry at a glance. The Company of Biologists; 2013.
- Gest H. The remarkable vision of Robert Hooke (1635–1703): first observer of the microbial world. *Perspect Biol Med*. 2005;48(2):266–72.
- LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436.
- Wiemken TL, Kelley RR. Machine learning in epidemiology and health outcomes research. *Annu Rev Public Health*. 2020;41:21–36.
- Brenner S, Horne R. A negative staining method for high resolution electron microscopy of viruses. *Biochim Biophys Acta*. 1959;34:103–10.
- LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE*. 1998;86(11):2278–324.
- Otsu N. A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern*. 1979;9(1): 62–6.
- Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer; 2015. p. 234–41.
- Gray RD, Beerli C, Pereira PM, Scherer KM, Samolej J, Bleck CKE, Mercer J, Henriques R. VirusMapper: open-source nanoscale mapping of viral architecture through super-resolution microscopy. *Sci Rep*. 2016;6:29132.
- Matuszewski DJ, Sintorn I-M. Reducing the u-net size for practical scenarios: virus recognition in electron microscopy images. *Comput Methods Prog Biomed*. 2019;178:31–9.
- Smith KP, Kang AD, Kirby JE. Automated interpretation of blood culture gram stains by use of a deep convolutional neural network. *J Clin Microbiol*. 2018;56(3):e01521–17.
- Jo Y, Park S, Jung J, Yoon J, Joo H, Kim M-H, Kang S-J, Choi MC, Lee SY, Park Y. Holographic deep learning for rapid optical screening of anthrax spores. *Sci Adv*. 2017;3(8):e1700606.
- Ho C-S, Jean N, Hogan CA, Blackmon L, Jeffrey SS, Holodniy M, Banaei N, Saleh AA, Ermon S, Dionne J. Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning. *Nat Commun*. 2019;10(1):1–8.
- Fisch D, Yakimovich A, Clough B, Wright J, Bunyan M, Howell M, Mercer J, Frickel E. Defining host-pathogen interactions employing an artificial intelligence workflow. *elife*. 2019;8:e40560.
- Yakimovich A, Huttunen M, Samolej J, Clough B, Yoshida N, Mostowy S, Frickel E-M, Mercer J. Mimicry embedding facilitates advanced neural network training for image-based pathogen detection. *Msphere*. 2020;5(5):e00836–20.
- Yakimovich A, Huttunen M, Samolej J, Clough B, Yoshida N, Mostowy S, Frickel E, Mercer J. Mimicry embedding for advanced neural network training of 3D biomedical micrographs. *bioRxiv*: 820076. 2019.
- Andriasyan V, Yakimovich A, Georgi F, Petkidis A, Witte R, Puntener D, Greber UF. Deep learning of virus infections reveals mechanics of lytic cells. *bioRxiv*:798074. 2019.
- Yakimovich A, Andriasyan V, Witte R, Wang I-H, Prasad V, Suomalainen M, Greber UF. Plaque2. 0 – a

- high-throughput analysis framework to score virus-cell transmission and clonal cell expansion. *PLoS One.* 2015;10(9):e0138760.
23. Ferrari A, Lombardi S, Signoroni A. Bacterial colony counting with convolutional neural networks in digital microbiology imaging. *Pattern Recogn.* 2017;61: 629–40.
24. Zieliński B, Plichta A, Misztal K, Spurek P, Brzyczy-Włoch M, Ochońska D. Deep learning approach to bacterial colony classification. *PLoS One.* 2017;12(9):e0184554.
25. Beerli C, Yakimovich A, Kilcher S, Reynoso GV, Fläschner G, Müller DJ, Hickman HD, Mercer J. Vaccinia virus hijacks EGFR signalling to enhance virus spread through rapid and directed infected cell motility. *Nat Microbiol.* 2019;4(2):216–225.
26. Müller V, Sousa JM, Koydemir HC, Veli M, Tseng D, Cerqueira L, Ozcan A, Azevedo NF, Westerlund F. Identification of pathogenic bacteria in complex samples using a smartphone based fluorescence microscope. *RSC Adv.* 2018;8(64):36493–502.
27. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, Askham H, Glorot X, O'Donoghue B, Visentin D. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med.* 2018;24(9):1342–50.
28. Gong E, Pauly JM, Wintermark M, Zaharchuk G. Deep learning enables reduced gadolinium dose for contrast-enhanced brain MRI. *J Magn Reson Imaging.* 2018;48(2):330–40.
29. Shi H, Han X, Jiang N, Cao Y, Alwalid O, Gu J, Fan Y, Zheng C. Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study. *Lancet Infect Dis.* 2020;20(4):425–434.
30. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med.* 2018;24(11):1716–20.
31. Laguarta J, Hueto F, Subirana B. COVID-19 artificial intelligence diagnosis using only cough recordings. *IEEE Open J Eng Med Biol.* 2020;1:275–281.
32. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. *Nat Genet.* 2019;51(1):12–8.
33. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol.* 2018;36(10):983–7.
34. Chiu CY, Miller SA. Clinical metagenomics. *Nat Rev Genet.* 2019;20(6):341.
35. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput.* 1997;9(8):1735–80.
36. Veltri D, Kamath U, Shehu A. Deep learning improves antimicrobial peptide recognition. *Bioinformatics.* 2018;34(16):2740–7.
37. Wei JN, Duvenaud D, Aspuru-Guzik A. Neural networks for the prediction of organic chemistry reactions. *ACS Cent Sci.* 2016;2(10):725–32.
38. Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, MacNair CR, French S, Carfrae LA, Bloom-Ackerman Z. A deep learning approach to antibiotic discovery. *Cell.* 2020;180(4):688–702.e613.
39. Beck BR, Shin B, Choi Y, Park S, Kang K. Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. *Comput Struct Biotechnol J.* 2020;18:784.
40. Yang JH, Wright SN, Hamblin M, McCloskey D, Alcantar MA, Schröbbers L, Lopatkin AJ, Satish S, Nili A, Palsson BO. A white-box machine learning approach for revealing antibiotic mechanisms of action. *Cell.* 2019;177(6):1649–1661.e1649.



Artificial Intelligence in Medicine: Modeling the Dynamics of Infectious Diseases

99

Richard Dybowski

Contents

Epidemiology	1379
History	1379
Using Artificial Intelligence	1380
The Recurrent Neural Network Approach	1380
The Multi-agent Approach	1381
Parameter Uncertainty	1384
Conclusion	1384
References	1384

Abstract

The classical approach to modeling, and thus predicting, the trajectories of outbreaks is to use the SIR model. This is a compartmental-based technique that has a number of limitations including limited population granularity. This chapter describes a richer agent-based approach. Computer requirements and the issue of parameter uncertainty are also discussed.

from neighborhoods to global. In this chapter, we will focus on the modeling of the dynamics of infectious diseases, particularly the use of AI for this purpose.

History

One reason for mathematically modeling an epidemic within a population is to attempt to predict how the epidemic will change over time. Following on from the initial observations made by Hamer [18] over 110 years ago, and the work of Ross and Hodges [33], Kermack and McKendrick [23] proposed that individuals can be classified according to their epidemiological status, namely, (i) those that are *susceptible* to the infection of interest, (ii) those that are *infected*, and therefore infectious, and (iii) those that have *recovered* and hence are no longer infectious. This SIR model is

Epidemiology

A broad definition of epidemiology is that it is the study of the distribution of health-related states and events (not just diseases), and the associated determinants, in specified populations ranging

R. Dybowski (✉)
St John's College, Cambridge, UK
e-mail: rd460@cam.ac.uk

based on the following set of ordinary differential equations:

$$\frac{dS}{dt} = -\alpha SI, \quad (1)$$

$$\frac{dI}{dt} = \alpha SI - \beta I, \quad (2)$$

$$\frac{dR}{dt} = \beta I, \quad (3)$$

$$S + I + R = N$$

where α is the disease transmission rate, β is the recovery (or death) rate, N is the total number of the whole population. $S = S(t)$, $I = I(t)$ and $R = R(t)$ are, respectively, the number of susceptible individuals, the number of infected individuals, and the number of recovered individuals at time t . See Fig. 1. The average number of infections arising initially from a single infected individual is the basic reproduction number $R_0 = \alpha/\beta$.

Although the SIR model provides a simple and generic framework for understanding and predicting epidemiological dynamics, a number of modifications are possible which increase model realism but also increase the number of parameters that have to be estimated (e.g., [15]).

SIR-type models assume population and spatial homogeneity, but this is often too simplistic an assumption. For example, motivated by the occurrence of measles amongst school children [35], Dietz and Schenzle [10] provided explicit expressions for the transmission potential of an

immunizing infection where the contact rates and the vaccination rates depend on the chronological age of an individual, and the infectivity and recovery rates depend on the duration of an infection. This consideration of the age structure of a population is also relevant in the context of the COVID-19 pandemic. One of the earliest considerations of the risk structure of a population was done with respect to sexuality heterogeneity during the HIV epidemic of the late 1980s and early 1990s [27].

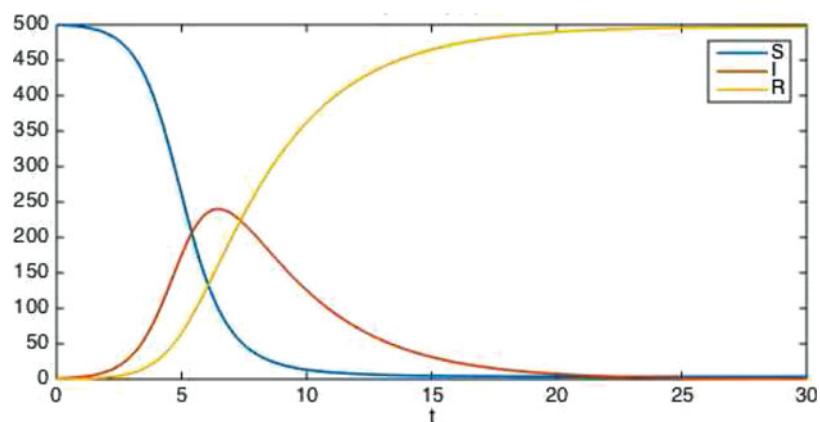
In addition to the assumption of uniformity of the spatial structure of an epidemic, SIR-type models also assume that populations are large; however, this assumption can be addressed by the use of stochastic models [2], an approach that has been used to model the spreading of influenza [25]. The foot-and-mouth disease epidemic of 2001 highlighted the importance of spatially explicit modeling as transmission between farms was a highly localized process [22].

Using Artificial Intelligence

The Recurrent Neural Network Approach

Recurrent neural networks (RNNs) [12, 21] with long short-term memory (LSTMs) [19] have performed impressively as time-series predictors, for example, in the realms of financial market prediction [44] and weather forecasting [5]. Therefore, it is natural to consider their use to predict the trajectories of epidemics.

Fig. 1 Typical plot of $S(t)$ (blue), $I(t)$ (red), and $R(t)$ (yellow) over time t



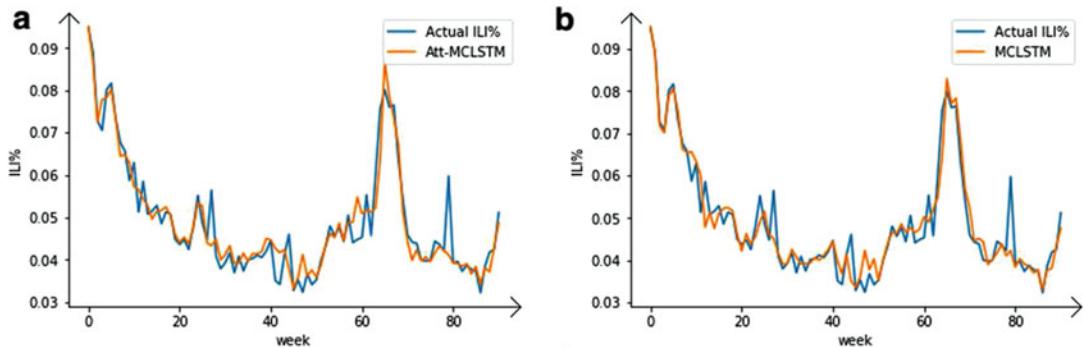


Fig. 2 Plot of observed influenza-like illness rate (ILI%) (blue) and ILI% predicted 1-week ahead (orange) using an RNN-LSTM (a) with attention mechanism, and (b) without attention mechanism [47]

Zhu et al. [47] applied a multi-input RNN with LSTM to predict the progress of the influenza-like illness rate 1 week ahead in Guangzhou, China (Fig. 2). They obtained a mean absolute percentage error (MAPE) of 0.118. When an attention mechanism [41] was added to the RNN-LSTM, the MAPE was reduced to 0.086.

Akhtar et al. [1] employed an RNN (but without LSTM) to predict the occurrence of the Zika virus across the Americas with an average 1-week-ahead accuracy of AUC 0.94. Inputs to the network included previous numbers of Zika cases and population densities.

With regard to forecasting the COVID-19 pandemic, Shahid et al. [36] showed that the use of a bidirectional LSTM gave better accuracy than a standard LSTM.

RNNs with LSTM generally have a good track record [40], but there is another AI-based technique that can produce rich epidemiological models, and it is to this approach that we turn next.

The Multi-agent Approach

Instead of using a set of differential equations (either deterministic or stochastic) to model the dynamics of an epidemic, an entirely alternative approach is to use an agent-based approach in which each member of a population is represented by an individual agent.

In general, a *multi-agent system* consists of multiple decision-making agents that interact in a shared environment to achieve common or

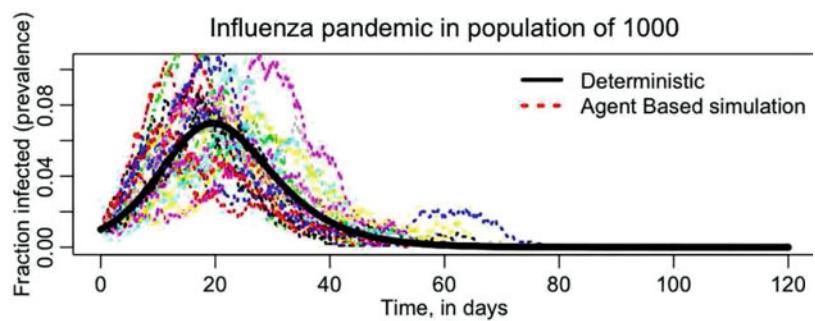
conflicting goals [37]. Multi-agent systems are regarded as a subfield of distributed artificial intelligence [45].

A general agent-based modeling scheme consists of the following steps. First, a set of A agents $\{a_1, \dots, a_A\}$ are initialized. Under this stage, agents are configured in a determined position or in a specific state. Then, each agent $a_i \in \{1, \dots, A\}$ is selected randomly or in a particular order. For a_i , a set of rules is applied in order to change its position, state, or relationship with other agents. These rules consider a relation of conditions imposed by other agents or local influences. This process is repeated until a predetermined stop criterion has been reached.

Agent-based models are generally simple – they do not use sophisticated architectures or difficult behavioral rules – but, in spite of these simple behaviors, they are capable of generating complex global patterns (behaviors) as a consequence of the interactions of a set of simple agents. Figure 3 shows a plot of $I(t)$ based on the deterministic SIR model (Eq. 2) for an influenza outbreak overlaid with 50 agent-based simulations of $I(t)$. Note the variety of plots obtained from the agent-based simulations. This stochasticity gives a more realistic measure of the uncertainty of $I(t)$.

The standard SIR-type epidemiological models are based on a number of assumptions that will not always be valid, for example, large and homogeneous populations. In contrast, agent-based epidemiological models are more flexible than the conventional equation-based compartmental

Fig. 3 Plot of $I(t)/N$ from Eq. 2 (black) for a hypothetical strain of influenza with $R_0 = 1.5$. The plot is overlaid with 50 agent-based simulations of $I(t)/N$ (various colors). $N = 1000$



models (e.g., [20]). They fully allow for heterogeneous populations, varied contact patterns, and relevant networks.

Furthermore, agent-based systems can be applied to small communities as well as to entire nations, and achieving such granularity allows interventions to be targeted locally. This would avoid unnecessary blanket lockdowns and their crippling economic and social costs.

Another advantage of agent-based systems is that the flow of infection over time within a community can be visualized. An example of this is the work of Hackl and Dubernet [17] that simulates the spread of an outbreak of influenza in the metropolitan area of Zurich over a 17-h period (Fig. 4).

Because agent-based approaches model individuals with distinct characteristics, they provide more realistic results. In particular, one can consider each agent's internal biology in the simulations. This would allow one to go beyond saying if someone is infected or not by being able to say that someone is infected and that they have a higher likelihood of transmission. This, in turn, allows a richer and potentially more accurate model of infection. An example of this is the work by Venkatraman et al. [42] on the use of an agent-based approach for forecasting the 2014–2015 Ebola epidemic in Liberia.

The above agent-based models were built for human-human transmissions, but what of vector-borne diseases such as malaria? This need was first approached by Roche et al. [32], and agent-based simulations have recently been proposed for schistosomiasis [8].

SARS-CoV-2

As regards the SARS-CoV-2 virus, Nanna et al. [29] developed a multi-agent system to model COVID-19 outbreaks. The human population, modeled as agents, behaved and interacted in a randomly generated urban setting, where the places with higher risk of contagion are considered. This included households, hospitals, public places, the transportation system, businesses, and mass gatherings. In their simulations, humans and businesses dynamically change their behavior depending on the measures and the restrictions enacted by the local government aimed at slowing down the spread of the disease. In order to simulate a more human behavior, they also admitted the possibility of reckless actions. Both Bouchnita and Jebrane [7] and Cuevas [9] provide technical details of the use of multi-agent systems for COVID-19 transmission simulations.

Magalhaes et al. [26] developed a simple multi-agent model that describes the transmission dynamics of coronavirus for a given location. This is done with respect to three parameters: population density; rate of social isolation rate; and the effective transmission probability represented by the Coronavirus Protection Index (CPI). The CPI is a measurement of a given territory's vulnerability to the coronavirus that includes characteristics of the health system, socioeconomic development as well as infrastructure. The model was calibrated by immunity surveys, and their simulations demonstrated the possible existence of multiple epidemic curves in the same city due to different vulnerabilities to the virus across regions.

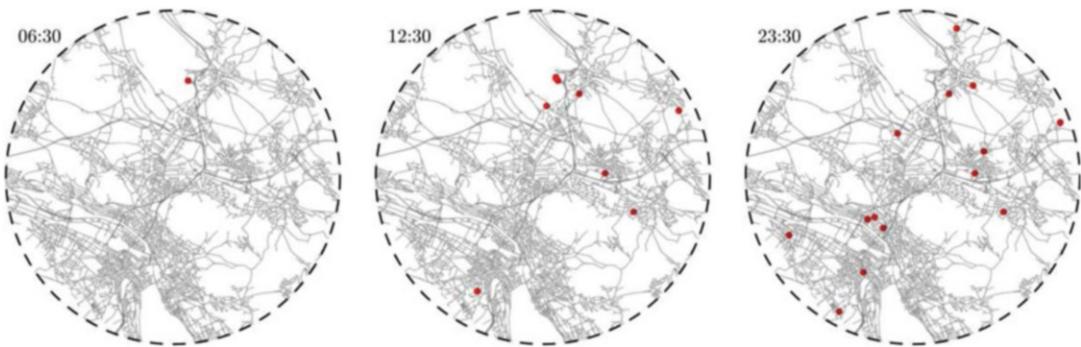


Fig. 4 Simulation of a temporal-spatial spread of influenza virus in a part of Zurich over 1 day starting with a single infected agent (red dot) at 6:30 am to many agents becoming infected by 11:30 pm [17]

The use of multi-agent systems to simulate and compare the efficacy of different outbreak intervention policies in silico has been adopted by a number of groups. For example, the model developed by Vyklyuk et al. [43] allowed them to compare the impact of various quarantine measures and restrictions on transport connections imposed by Ukraine, Slovakia, Turkey, and Serbia. Silva et al. [38] also compared intervention policies looking at economic impacts as well as consequences on health.

Fiore et al. [13] used multi-agent simulations to estimate the testing capacity required to find and isolate a number of infections sufficient to break the chain of transmission of SARS-CoV-2. Depending on the mitigation policies in place, a daily capacity between 0.7 and 3.6 tests per thousand was required to contain the disease. However, if contact tracing and testing efficacy dropped below 60%, the number of infections kept growing exponentially, irrespective of any testing capacity.

Contact tracing has become an important part of the attempt to manage the COVID-19 pandemic, and Omae et al. [30] used a multi-agent system to simulate the effectiveness of a Japanese contact-tracing app in various scenarios.

The effectiveness of vaccination campaigns was investigated by Nadini et al. [28], who supported the intuition that vaccination in central and dense areas urban should be prioritized.

During the early phase of the pandemic, *The Lancet* wrote that there should be research to “determine the best ways to apply knowledge about infection prevention and control in healthcare settings in resource-constrained countries” [4]. Furthermore, an article by *the Economist* highlighted how the pandemic is likely to devastate poor countries with healthcare systems that are not in a position to cope [11]. This need to support poor countries during the pandemic is still of immense importance, and this has led Gilman et al. [14] and others to focus on the use of multi-agent systems to determine the optimal outbreak interventions to be used in refugee camps.

There have been other AI applications in response to the coronavirus pandemic [3], such as predicting the impact of air travel [6], early diagnosis of COVID-19 via CT scans [39], CoV-2 protein structure prediction [31], and robots for COVID-19 field hospitals [34], but here our focus has been on infection dynamics.

Computing Requirements

Computational efficiency is an important concern when modeling the potential interactions of millions of individuals. Although multi-agent systems are far more flexible than the conventional deterministic models, they are also far more computationally expensive and can require fast, parallel processors.

Various software packages have been developed to facilitate the construction of multi-agent

systems, including IDESS [46] and FRED [16]. FRED (a Framework for Reconstructing Epidemic Diseases) is a freely available open-source epidemic modeling system that uses census-based synthetic populations to capture the demographic and geographic heterogeneities of the population, including realistic household, school, and workplace social networks. FRED requires between 750 and 1000 MB of memory per million simulated individuals. The exact amount of memory required depends on the demographic and geographic characteristics of the synthetic population, as well as the severity of the simulated epidemic. Simulations of an influenza spread like the H₁N₁ pandemic in a population of 1 million people takes less than 2 min on a typical dual-core laptop computer but the runtime will vary depending on the number of individuals infected during the epidemic. For example, on an SGI Altix UV shared-memory architecture with up to 16 TB of shared memory, a simulated pandemic over the entire United States population requires approximately 200 GB of memory and takes approximately 4 h using 16 threads.

Born out of the games industry, which demands high fidelity simulations, a recent development is the Aether Engine produced by Hadean [24], which can be readily sealed up to provide large-scale simulations of COVID-19 outbreaks within the UK and beyond.

Parameter Uncertainty

The plots shown in Fig. 1 were based on fixed (i.e., known) values for parameters α and β in Eqs. 2 to 3; however, in reality, these are estimates with an associated uncertainty.

Suppose that, in place of a single value for α , we have a probability distribution function $\mathbb{P}(\alpha)$ that expresses the uncertainty in α . Similarly, for β , we have a probability distribution function $\mathbb{P}(\beta)$. More honest plots for $S(t)$, $I(t)$, and $R(t)$ could then be obtained by randomly selecting α and β from $\mathbb{P}(\alpha)$ and $\mathbb{P}(\beta)$, respectively, obtaining the resulting plot, and then repeating this a large number of times, each time randomly selecting α and β .

In the case of multi-agent systems, the same approach could be adopted for all relevant parameters. Note that this would give an additional stochasticity on top of the intrinsic stochasticity resulting from using multi-agent systems. But these uncertainties lead to the important question of how best to respond to epidemics under (sometimes extreme) uncertainty.

Conclusion

Recurrent neural networks have significantly improved forecasting over traditional approaches to time-series modeling, such as the autoregressive integrated moving average (ARIMA) [36].

As discussed, the multi-agent approach provides a richer modeling of infectious disease outbreaks than the compartmental models. These are visual over a geospatial display and are not restricted to populations that are large and homogeneous. Furthermore, the availability of powerful computers with efficient algorithms means that multi-agent systems are scalable and thus a realistic alternative to the conventional forecasting techniques.

References

1. Akhtar M, Kraemer M, Gardner L. A dynamic neural network model for predicting risk of Zika in real time. *BMC Med.* 2019;17:171.
2. Allen L. A primer on stochastic epidemic models: formulation, numerical simulation, and analysis. *Infect Dis Model.* 2017;2:128–42.
3. Arora N, Banerjee A, Narasu M. The role of artificial intelligence in tackling COVID-19. *Future Virol.* 2020. <https://doi.org/10.2217/fvl-2020-0130>.
4. Bedford J, Enria D, Giesecke J, Heymann D, Ihkweazu C, Kobinger G ... Wieler L. COVID-19: towards controlling of a pandemic. *Lancet.* 2020;395(10229):1015–8.
5. Biswas S, Sinha N, Purkayastha B, Marbaniang L. Weather prediction by recurrent neural network dynamics. *Int J Intell Eng Inform.* 2014;2(2/3):166–80.
6. Bogoch I, Watts A, Thomas-Bachli A, Huber C, Kraemer M, Khan K. Pneumonia of unknown aetiology in Wuhan, China: potential for international spread via commercial air travel. *J Travel Med.* 2020;27(2):taaa008.

7. Bouchnita A, Jebrane A. A hybrid multi-scale model of COVID-19 transmission dynamics to assess the potential of non-pharmaceutical interventions. *Chaos Solitons Fractals*. 2020;138:109941.
8. Cisse P, Dembele J, Lo M, Cambier C. Multi-agent systems for epidemiology: example of an agent-based simulation platform for schistosomiasis. In: Montagna S, Abreu P, Giroux S, Schumacher M, editors. A2HC: international workshop on agents applied in health care. AHEALTH: international workshop on agents and multi-agent systems for AAL and e-health. Porto; 2017. Springer, p. 131–53.
9. Cuevas E. An agent-based model to evaluate the COVID-19 transmission risks in facilities. *Comput Biol Med*. 2020;121(103827).
10. Dietz K, Schenzle D. Proportionate mixing models for age-dependent infection transmission. *J Math Biol*. 1985;22:117–20.
11. Economist. The coronavirus could devastate poor countries. 2020, March. https://warwick.ac.uk/fac/soc/economics/staff/amuthoo/pandemics/impact_on_poor_countries_from_the_economist_26_march_2020.pdf. Accessed 26 Mar 2020.
12. Elman J. Finding structure in time. *Cogn Sci*. 1990;14(2):179–211.
13. Fiore V, DeFelice N, Glicksberg B, Perl O, Shuster A, Kulkarni K ... Gu X. Containment of future waves of COVID-19: simulating the impact of different policies and testing capacities for contact tracing, testing, and isolation. *medRxiv*. 2020. <https://doi.org/10.1101/2020.06.05.20123372>.
14. Gilman RT, Mahroof-Shaffi S, Harkensee C, Chamberlain A. Modelling interventions to control COVID-19 outbreaks in a refugee camp. *BMJ Glob Health*. 2020;5:e003727.
15. Giordano G, Blanchini F, Bruno R, Colaneri P, Di Filippo A, Di Matteo A, Colaneri M. Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nat Med*. 2020;26:855–60.
16. Grefenstette, J., Brown, S., Rosenfeld, R., DePasse, J., Stone, N., Cooley, P., ... Burke, D. FRED (A Framework for Reconstructing Epidemic Dynamics): an open-source software system for modeling infectious diseases and control strategies using census-based populations. *BMC Public Health*. 2013;13(9):140.
17. Hackl J, Dubernet T. Epidemic spreading in urban areas using agent-based transportation models. *Future Internet*. 2019;11:92.
18. Hamer W. Epidemic diseases in England: the evidence of variability and of persistency of type. *Lancet*. 1906;1:733–9.
19. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9(8):1735–80.
20. Hunter E, MacNamee B, Kelleher J. A taxonomy for agent-based models in human infectious disease epidemiology. *J Artif Soc Soc Simul*. 2017;20(3):2.
21. Jordan M. Serial order: a parallel distributed processing approach. *Adv Psychol*. 1997;121:471–95.
22. Keeling M, Woolhouse M, Shaw D, Matthews L, Chase-Topping M, Haydon D ... Grenfell B. Dynamics of the 2001 UK foot-and-mouth epidemic: stochastic dispersal in a heterogeneous landscape. *Science*. 2001;294:813–7.
23. Kermack W, McKendrick A. Contribution to the mathematical theory of epidemics. *Proc R Soc A*. 1927;115: 700–21.
24. Keshani M. Contagion modelling: applying spatial simulation to track pathogen spread. 2020. <https://hadean.com/blog/contagion-modelling-applying-spatial-simulation-to-track-the-spread-of-pathogens/>. Accessed 24 Oct 2020.
25. Liccardo A, Fierro A. A lattice model for influenza spreading. *PLoS One*. 2013;8(5):e63935.
26. Magalhaes P, Pinto J, Angel DMS. A multiagent coronavirus model with territorial vulnerability parameters. *medRxiv*. 2020. <https://doi.org/10.1101/2020.10.25.20218735>.
27. May R, Anderson R. Transmission dynamics of HIV-infection. *Nature*. 1987;326:137–42.
28. Nadini M, Zino L, Rizzo A, Porfiri M. A multi-agent model to study epidemic spreading and vaccination strategies in an urban-like environment. *Appl Netw Sci*. 2020;1:68.
29. Nanna G, Quatraro N, De Caroli B. A multi-agent system for simulating the spread of a contagious disease. In: Woa 2020: workshop “from objects to agents”. Bologna; 2020. CEUR, p. 119–34.
30. Omae Y, Toyotani J, Hara K, Gon Y, Takahashi H. Effectiveness of the COVID-19 contact-confirming application (COCOA) based on a multi-agent simulation. *arXiv*. 2020. 2006.13166.
31. Pfab J, Phan N, Si D. DeepTracer for fast de novo Cryo-EM protein structure modeling and special studies on CoV-related complexes. *PNAS*. 2021;118(2): e2017525118.
32. Roche B, Guégan J-F, Bousquet F. Multi-agent systems in epidemiology: a first step for computational biology in the study of vector-borne disease transmission. *BMC Bioinformat*. 2008;9:435.
33. Ross R, Hodges H. An application of the theory of probabilities to the study of a priori pathometry – part III. *Proc R Soc A*. 1917;93:225–40.
34. Sayed A, Ammar H, Shalaby R. Centralized multi-agent mobile robots SLAM and navigation for COVID-19 field hospitals. In: Proceedings of NILES 2020: 2nd novel intelligent and leading emerging sciences conference. Giza: IEEE Press; 2020. p. 444–9.
35. Schenzle D. An age-structured model of pre- and post-vaccination measles transmission. *IMA J Math Appl Med Biol*. 1984;1(2):169–91.
36. Shahid F, Zameer A, Muneeb M. Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos Solitons Fractals*. 2020;140:10212.

37. Shoham Y, Leyton-Brown K. Multiagent systems. Cambridge: Cambridge University Press; 2009.
38. Silva P, Batista P, Lima H, Alves M, Guimaraes F, Silva R. COVID-ABS: an agent-based model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions. *Chaos Solitons Fractals*. 2020;139(110088).
39. Singh D, Kumar V, Vaishali K. Classification of COVID-19 patients from chest CT images using multi-objective differential evolution-based convolutional neural networks. *Eur J Clin Microbiol Infect Dis*. 2020;39(7):1379–89.
40. Tealab A. Time series forecasting using artificial neural networks methodologies: a systematic review. *Future Comput Informat J*. 2018;3:334–40.
41. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, ... Polosukhin I. Attention is all you need. *arXiv*. 2017. 1706.03762.
42. Venkatraman S, Lewis B, Chen J, Higdon D, Vullikanti A, Marathe M. Using data-driven agent-based models for forecasting emerging infectious diseases. *Epidemics*. 2018;22:43–9.
43. Vyklyuk Y, Manylich M, Skoda M, Radovanovic M, Petrović MD. Modeling and analysis of different scenarios for the spread of COVID-19 by using the modified multi-agent systems evidence from the selected countries. *Results Phys*. 2021;20: 103662.
44. Wang Y. Applications of recurrent neural network on financial time series (Unpublished master's thesis). Imperial College. 2017.
45. Weiss G, editor. Multiagent systems: a modern approach to distributed artificial intelligence. Cambridge, MA: MIT Press; 1999.
46. Yergens D, Hiner J, Denzinger J, Noseworthy T. Multi agent simulation system for rapidly developing infectious disease models in developing countries. 2006. https://pages.cpsc.ucalgary.ca/~denzinge/papers/AAMASHealth_Yergens_Final.pdf. Accessed 02 Feb 2021.
47. Zhu X, Fu B, Yang Y, Ma Y, Hao J, Chen S, ... Liao Z. Attention-based recurrent neural network for influenza epidemic prediction. *BMC Bioinformat*. 2019;20:575.



Arash Keshavarzi Arshadi and Milad Salem

Contents

Introduction	1388
AI and Vaccine Discovery	1389
AI and Vaccine Discovery for SARS-CoV-2	1390
AI for MHC Binding Peptide Prediction and Organ Transplant	1390
The Challenge of Variation	1391
Drug Discovery and Vaccine Discovery Differences for Learning	1391
Discussion	1392
References	1393

Abstract

In the last decade, artificial intelligence (AI) has revolutionized many scientific areas. Due to huge amount of data being generated in biomedical fields, AI has been a crucial computational tool to improve predicting modeling and simulation. AI can find hidden pattern in biological data (genomic, proteomic, transcriptomic, small molecules, etc.) and connect the unknown parts of a big puzzle. Immune system is for sure one of the most complicated molecular networks in human body. A good question would be how these immune cells and

proteins recognize the normal inner cells from other foreign or abnormal cells (infected or cancerous cells) and viruses. For many years, scientists have been working to answer this question in detail. Understanding the immune system is equal to ability to trigger it against strong viruses and bacteria. AI has been helping scientists in the field immunoinformatic and vaccine discovery to improve the efficacy of predicted epitopes. However, dealing with immunological data has been very challenging, and the representation of each sample has not been easy. In this chapter, we discuss how AI started to improve vaccinology and how it will be leading the future of computational vaccine development and immunoinformatic.

A. Keshavarzi Arshadi (✉) · M. Salem
University of Central Florida, Orlando, FL, USA
e-mail: arashka@knights.ucf.edu;
miladsalem@knights.ucf.edu

Keywords

Artificial intelligence · Deep learning · Machine learning · Immunology · Vaccinology · HLA typing · Transplantation

Introduction

Immunology is for sure one of the most crucial subdivision in biomedical sciences. After an infection in body, a very complex and organized system of cells and their secreted proteins start many complicated processes in order to overcome the foreign attach and keep a record of it for future references. However, this system would also recognize and destroy any inner threat such as cells out of their natural cycles including cancerous ones [1]. In evolution, a very smart way of distinguishing cells coded in our DNA has been developed. In a simple language, a very important genomic complex entitled major histocompatibility complex (MHC) is responsible for a system of critical proteins named human leukocyte antigen (HLA) [2–4]. These genes are very polymorphic and can vary between different people. Any infected cell would just present fragments of the bacteria or viruses' proteins, epitopes, on surface by HLA system and then the immune system would do the rest and get rid of that cell [2].

The idea of human body having an inner immune system goes back to thousands of years ago. One of the first mentions in this regard is recorded in Gilgamesh poem, a written document in Babylonian period (2000 BC) [5]. Thucydides was an Athenian historian who understood some aspects of acquiring immunity from prior infection in plague era of Athens (430 BC) [6]. In a revolutionary breakthrough of immunology history, two Iranian (Persian) alchemist and philosophers, Rhazes in tenth century and Avicenna in eleventh century, developed the idea of adoptive immune system from understanding the fact that fever is an immune process to combat diseases [7]. Avicenna, who is called the father of modern medicine by many, further developed the idea of acquired immunity and made a revolution in this regard [8, 9]. In addition, one of the first practical

implementation of vaccines and intriguing of adaptive immune system was done by Edward Jenner, an English physician who used crusts from smallpox pustules as the world's first vaccine against smallpox [10]. Similar methods are reported to be executed in 1000 AD in China. In 1880, Louis Pasteur and Emile Roux attenuated cultures in chicken which lead to the induction of immune system and a very first artificially made vaccine. Nowadays, vaccines are widely practiced in the whole world for the artificial induction of immune system (*Vaccines and Immunization, WHO*).

Humans mainly have two central types of immune response, innate and adoptive immune system. Innate immune system is basically older in evolution as the first line of defense in most of creatures including insects, mammalians, and fungi. One of the primary functions of this system is informing and recruiting another immune components using some chemicals called cytokines [11]. The second immune response which is smarter and more specialized in the recognition of already encountered pathogens is adoptive immune system. This immune system is seen in vertebrates, and it consists of both humoral and cell-based immunity [12, 13]. Humoral immune response consists of mostly macromolecules such as antimicrobial peptides, antibodies, and defensins. On the other hand, cell-based immunity includes the cell-mediated responses to pathogens, as an example we can point to the cytokine secretion or macrophage involvement. The biggest distinguishable difference between the innate and adoptive is the specialized reaction [14]. Adoptive immunity is fundamentally a specialized memory response to pathogens. This memory would last months to couple of years as protection and in some cases a lifelong immunity. The idea of vaccination and artificial immunity is literally generated by understanding the fact that immunity's memory and response can last longer than the infection duration. Epitopes' generation and representation to immune system is the core approach for mimicking the natural long-term immunity memory [15].

Old school vaccine discovery and development is very tedious, time consuming, and

expensive. In these approaches, scientists have to culture the pathogens (for bacterial vaccine discovery) or grow cells containing pathogenic particle (for virus vaccine discovery), inactivate them, and inject them to healthy and nonhealthy volunteers [16]. This may take years for vaccine production, testing, and trial. Therefore, there has been an urgent need for alternative approaches and technologies for vaccine discovery [17]. After the introduction of different biological assays to measure diverse factors of immune responses and other variables, huge datasets started to pile up in many laboratories of universities and companies. Soon a new concept was born, computational analysis of big immunological datasets or immunoinformatic. With the rise of robots and high throughput assays, there has been a pleading need for new software and algorithms to analyze the large, generated datasets [18].

Basic areas of immunoinformatics include T cell and B cell epitope prediction, in silico vaccine discovery, discovering immune-related genes, allergy prediction, virulence gene identification, and most importantly, HLA typing and sequencing, database creation, and design [19–21]. Old school modeling and algorithm implementation has not been able to leverage the full potential of huge databases created by many diverse immunological assays. Also, the accuracy of those modeling has not been satisfying. Most of the time, the hidden patterns and information inside the mentioned databases are way more complicated for scientists to discover and distinguish. As an example, we can mention the HLA typing process. In spite of 8/8 HLA matching, many patients have complications after transplantation (we will talk more about HLA typing in following sections). Therefore, there should a type of modeling and algorithm capable of learning hidden pattern in an abstract way. Machine learning comes to rescue!

AI and Vaccine Discovery

With the emergence of artificial intelligence in last decades, automation has become a new reality in many industries. From autonomous driving and credit card fraud detection to face recognition and

cancer detection, artificial intelligence has revolutionized many fields [22–26]. Biomedical sciences inherently are a type of field that deals with very large amounts of generated data. From hospitals to biological labs, a huge number of genomic, molecular, proteomic, etc. class of data is collected every day. Old school type of data analysis and modeling has not been capable of getting the most out of biomedical big data. Therefore, there has been a need for newer computational technologies for analysis and prediction. Artificial intelligence comes to rescue. AI helps the scientists gain insights via combining multiple datasets from real-world cases which are inherently messy. In vaccine discovery, AI aids in the same manner and can help with recognizing which pieces of the virus are more likely to be recognized by the immune system [27], which epitopes are likely to be presented on the surface of the cells [28], and which peptides can be toxic.

The search for a vaccine has four steps, finding the vaccine's target, determining the needed immune response, generating the needed response, and discovering what responses are generated in the vaccine's subjects. The first two steps are more understood than the latter steps. AI and modeling can help make sense of the complexity of the immune system, accelerate the process of vaccine discovery, and increase the probability of vaccine's success [19, 29]. System Vaccinology tries to predict the beneficial and harmful responses generated from the vaccine in the forms of immunogenicity and reactogenicity [30]. AI can be used in this regard to learn from different input data sources such as clinical, genome, transcriptome, proteome, metabolome, mass cytometry, and cytokine profiling data sources to model and predict outputs such as antibody response, T cell response, adverse events, and vaccine protection [31].

Another area which has been impacted by the use of AI is vaccine distribution and planning. Traditional machine learning models such as Random Forest (RF) and Support Vector Machine (SVM) have been used to identify children who may drop out from the immunization course and not follow up on the subsequent immunization visit [32]. These predictions can be used to better

plan for these high-risk individuals and ensure universal immunization coverage.

Use of AI in vaccine discovery has also faced challenges. One of the main challenges for AI is that it cannot replace one of the most time-consuming steps in the vaccine discovery process, which is animal and human trials (Emily Waltz, 2020). Moreover, imbalanced datasets with numerous features which are prevalent in the field of immunology still pose a challenge to machine learning algorithms [31].

AI and Vaccine Discovery for SARS-CoV-2

Due to the nature of the pandemic and the necessity of a prompt response, AI-based approaches were heavily used by researchers around the world to accelerate the vaccine discovery process for SARS-CoV-2 [33]. In the early days of the outbreak, the Stanford Institute for Human-Centered Artificial Intelligence (HAI) used neural network-based tools such as NetMHCpan 4.0 [28], MARIA [32], and linear regression-based tool DiscoTope 2 to release a list of possible epitopes for SARS-CoV-2 [35].

Deep-Vac-Pred trains a convolutional neural network to aid in the design of a multi-epitope vaccine for SARS-CoV-2. This model predicts potential vaccine subunits from the virus's spike protein sequence and filters the predictions via investigating the linear B cell epitopes, cytotoxic T lymphocytes (CTL) epitopes, and the helper T lymphocytes (HTL) epitopes [36]. An innovation of this work is the dataset design, in which the positive data points must contain at least one B cell epitope, one T cell epitope, and be protective antigens. This dataset design narrows the predictions immensely and accelerates the process overall. Further in silico tools are utilized to filter the prediction, i.e., BepiPred [37], NetMHCpan 4.1 [28], and NetMHCIIPan [27], and to evaluate the quality of the designed vaccine, i.e., Vaxijen 2.0 [38] and ToxinPred [39].

AI has also been a useful resource in predicting the 3D structure of several proteins linked to SARS-CoV-2. The structure of the protein plays

an important role in defining how the protein functions, and in the case of SARS-CoV-2, how the protein can be targeted. Therefore, many researchers used different tools, including AI and deep learning, to predict these 3D structures. DeepMind provided these predictions using AlphaFold [40], a convolutional neural network-based model which learns to predict the pairwise distances between the amino acids of the protein sequence, then optimizes the potential of mean force to find the shape of the protein. Another model which assisted the prediction of SARS-CoV-2 protein structures was trRosetta [41]. This model consisted of a deep residual convolutional neural network which learns to predict the orientation of all residue pairs alongside their relative distance. Some of the main predictions made by AlphaFold and trRosetta were soon after verified to be accurate [42].

AI for MHC Binding Peptide Prediction and Organ Transplant

It has been shown that for having a recognition of peptides by T cells, they need to bind to MHC molecules first. As mentioned before, major histocompatibility complex (MHC) are a class of cell surface proteins (also called human leukemia antigen or HLA) that are coded by a very polymorphic region of DNA. One of the most important function of HLA molecules is the regulation of immune system [2, 43]. These molecules also have the responsibility of distinguishing between self and outside proteins, therefore they can be very crucial in organ transplantation. Any mutation in these genes would be destructive and result in autoimmune-related problems. As a rule of thumb, any cell representing a nonfamiliar HLA type is considered a threat by immune system and will be removed by it. Transplant rejection is basically a reason of representing the non-self HLA by the new organ [44].

HLA typing is an important step before organ donor selection. We inherit these biomarkers from our both parents. That is why the chance of finding a perfect HLA match in our family is higher than unrelated people [45, 46]. Also, it is more

probable that we can find a match in the same race of the patients than farther ones. Nowadays, health care organizations and hospitals suggest matching at least 6 out of 8 loci of HLAs to decrease the chance of having problems after transplantation [47]. However, still about 40% of patients experience complications and death after 8 of 8 matching. To find a solution for this problem, artificial intelligence and immunoinformatic come to rescue [48]. AI can find unknown loci and other important biomarkers by discovering the hidden patterns in immunological data.

The Challenge of Variation

Class I HLAs commonly bind to short peptides which are around nine-mers; however, Class II HLAs usually have the capability to bind to larger peptides around 15-mers and longer. It has been theorized that even though these peptides are longer, the binding core is still around nine-mers similar to Class I HLAs [49]. The variation in binding peptide length for both classes becomes a challenge for machine learning methods which use fix-sized input length, such as fully connected neural networks used in NetMHCpan or NetMHCIIPan [28], nHLAPred [50], and IEDB [51]. This challenge necessitates first detecting the binding core within the longer peptides, which requires preprocessing steps.

Aside from variation in peptide length, the diversity within the HLAs impacts binding, presentation, and the immune response. To accommodate this variation, two general methods have been developed. First approach is to train different models for different alleles, which are usually selected based on the amount of available data or the importance of the given allele. The second approach implemented in response to the variation within the HLAs is to train a single pan-allelic model which can make predictions unrestricted by the input allele. Commonly a pseudo-sequence is used to represent the allele and normalize the input. Removing the allele restriction and training a pan-allelic model ensures that the model encounters different alleles during training and can learn generalizable and sharable features for

the alleles which may increase the prediction performance of the model.

However, pan-allelic models may still be challenged if the input HLA is rare with scarce training data. This shortcoming is addressed in works such as MixMHCp [52] and RBM-MHC [53] which use semi-supervised approaches to generate candidate presentable peptides for rare HLAs. Rare alleles also present a challenge for genotyping given their low frequency. DEEP*HLA [54] alleviates this problem by using convolutional neural networks for allelic imputation.

Drug Discovery and Vaccine Discovery Differences for Learning

As mentioned before, artificial intelligence has proved its strength in automizing many fields. However, because of inherent differences in different fields, AI has been implemented in different ways. Therefore, the challenges in these fields can be diverse as well. In computer vision, data is abundant, and often easily labeled in a costly manner. One other such example would be structured hospital-driven data which would be list of patients with different features and characteristics like the type of disease, severity of the disease, the age of patient, etc. However, some data are inherently not structured, like molecular data. Each molecule can have different creative shapes in space due to diverse characteristics of atoms to form various types of bonds and angles. Having that said, structured representations would not be helpful for such type of data. Therefore, different types of modeling and data representation would be crucial to train a learning-based model [55].

Graph convolutional neural network (GCNN) are a suitable candidate for the unstructured case of drug discovery. Each molecule can be presented as a graph, with nodes representing atoms and edges representing the bonds between the atoms. Therefore, a model capable of learning graph-based data would be beneficial for drug discovery field. GCNN has proved that it can reach amazing results for drug discovery, and there are multiple helpful resources to implement these types of libraries like DeepChem

[25, 56]. But what about vaccine discovery? Are antibodies and epitopes representation the same as simple small molecules? Is genomic sequence of HLA loci the same as well? These are some important questions we should investigate here.

Even though antibodies or epitopes are basically molecules, but they are not small. Proteins are often bigger in size, and they are made of small molecular compartments called amino acid. Therefore, proteins can simply be viewed a sequence of at least 20 different types of small molecules. Since antibodies and epitopes are proteins too, we should consider other types of representation and modeling for them. This is also the case for DNA or RNA as well which are sequenced based; however, the length of these sequences is even longer which presents a higher challenge. A normal DNA consists of four compartments called adenine guanine cytosine and thymine (AGCT) [57].

One example of such long sequences is HLA loci, which is about 3 Mbp in length. However, protein sequences are often much shorter in length. For example, anti-microbial peptides (AMP) which are among the first defensive lines against microbes and even cancerous cells are 10–50 base pare in length [58]. If we want to compare the input data of a normal small molecules data to a variant call format (VCF) data of different SNPs as samples, we should mention that a VCF data can have millions of different features in itself, but a small molecular data would have tens or hundreds of features and label in total. Therefore, dealing with molecular data is easier in this regard. However, small molecular data from a phenotypic screening assay also have the challenge of being highly imbalanced. A problem that would not be the case for genomic or proteomics data. Balancing the low number of active small molecules has been one big challenge for data scientists and cheminformatician so far and some methods like transfer learning has been applied to overcome the challenge [59].

Overall, in the case of molecules, GCNNs have proven to be useful while in the case of sequential input, sequence-based models such as recurrent neural networks are often successful in learning representations. One main point that unites these two fields together is the benefits of deep learning.

The traditional approaches in these fields, similar to many other fields, are extracting subject matter expertise related features from the data as an input to a simple machine learning model. While these traditional approaches are tried and true, they often lack in generalization and efficiently learning from a bigger size of input data.

However, deep learning solves these challenges by combining the feature extraction and modeling stages together, and automatically extracting features which are helpful for modeling during training. This automatic representation learning can be viewed as the main advantage of deep learning and might result in predictions or insights that were previously overlooked or undiscovered from the perspective of subject matter experts. The same as medicinal chemist being unable to detect all molecular features, biochemists are not able to detect all hidden complicated features in a 3D representation of proteins. As an example, we should investigate antibodies more which are very complex proteins with some reserved and diverse compartments. In most of the cases, it is better to let the deep learning model to discover hidden patterns in all first, secondary, and tertiary formations. However, deep learning models are not that easy to interpret what is happening when the model is training and featurizing the data [60].

Discussion

Artificial intelligence is the new reality of automation. Considering the big data being generated in biomedical fields every day, AI has been one of the best computational tools to leverage the valuable information inside. Pandemics has shown us that old school vaccine discovery are not fast enough to save millions of people, and there should be cheaper, faster, and more efficient tools to discover and approve vaccines. In addition, the huge percentage of patience experiencing complication after HLA matching is another proof that newer method is in need for discovering unknown biomarkers important in immune response. One crucial part of any vaccine development is clinical trial. The trial would take many years and it is for sure very tedious and expensive.

One of the biggest challenges for AI in future would be solving the problem of human trial. The goal would be having a fast model capable of predicting the chance of any vaccine's failure in clinical trial. This way, vaccines can get developed and approved in a very fast pace, helping societies allocate their resources in a more efficient manner. These challenges appear solvable via discovering newer and better technologies for data representation, modeling, and learning.

References

- Warrington R, Watson W, Kim HL, Antonetti FR. An introduction to immunology and immunopathology. *Allergy Asthma Clin Immunol*. 2011;7(S1):S1. Available from: <https://aacijournal.biomedcentral.com/articles/10.1186/1710-1492-7-S1-S1>
- Wieczorek M, Abualrous ET, Sticht J, Álvaro-Benito M, Stolzenberg S, Noé F, et al. Major histocompatibility complex (MHC) class I and MHC class II proteins: conformational plasticity in antigen presentation [Internet]. *Front Immunol*. 2017;8:1. Frontiers Research Foundation. Available from: www.frontiersin.org
- Berger A. HLA typing. *BMJ*. 2001;322(7280):218. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1119473/>
- Charles A Janeway J, Travers P, Walport M, Shlomchik MJ. The major histocompatibility complex and its functions. 2001 [cited 2021 Feb 16]. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK27156/>
- Tarantola A. Four thousand years of concepts relating to rabies in animals and humans, its prevention and its cure [Internet]. *Trop Med Infect Dis*. 2017;2, MDPI AG, [cited 2021 Feb 16]. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6082082/>
- Littman RJ. The plague of Athens: epidemiology and paleopathology [internet]. *Mt Sinai J Med*. 2009;76: 456–67. Available from: <https://pubmed.ncbi.nlm.nih.gov/19787658/>
- Band IC, Reichel M. Al rhazes and the beginning of the end of smallpox [internet]. *JAMA Dermatol*. 2017;153:420. American Medical Association. Available from: <https://pubmed.ncbi.nlm.nih.gov/28492840/>
- Doherty M, Robertson MJ. Some early trends in immunology. *Trends Immunol*. 2004;25:623–31. Elsevier Current Trends.
- Dalfardi B, Esnaashary MH, Yarmohammadi H. Rabies in medieval Persian literature – the Canon of Avicenna (980–1037 AD). *Infect Dis Poverty*. 2014;3(1):7. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3933285/>
- Riedel S. Edward Jenner and the History of Smallpox and Vaccination. *Baylor Univ Med Cent Proc*. 2005;18(1):21–5. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1200696/>
- Dueque GA, Descoteaux A. Macrophage cytokines: involvement in immunity and infectious diseases [Internet]. *Front Immunol*. 2014;5:491. Frontiers Media S.A.; [cited 2021 Feb 16]. Available from: www.frontiersin.org
- Weinberger SE, Cockrill BA, Mandel J. Lung defense mechanisms. In: *Principles of pulmonary medicine*. Elsevier: Amsterdam, Netherlands 2019. p. 285–96.
- Harasymowicz NS, Rashidi N, Savadipour A, Wu C, Tang R, Bramley J, et al. Single-cell RNA sequencing reveals the induction of novel myeloid and myeloid-associated cell populations in visceral fat with long-term obesity. *FASEB J*. 2021;35(3):e21417. Available from: <https://onlinelibrary.wiley.com/doi/10.1096/fj.202001970R>
- Hoebe K, Janssen E, Beutler B. The interface between innate and adaptive immunity [Internet]. *Nat Immunol*. 2004;5:971–4. Nature Publishing Group. Available from: <http://www.nature.com/natureimmunology>
- Cooke SN, Ovsyannikova IG, Kennedy RB, Poland GA. Immunoinformatic identification of B cell and T cell epitopes in the SARS-CoV-2 proteome. *Sci Rep*. 2020;10(1):14179. <https://doi.org/10.1038/s41598-020-70864-8>.
- Plotkin SA. Vaccines, vaccination, and vaccinology [Internet]. *J Infect Dis*. 2003;187:1349–59. Oxford Academic. Available from: <https://academic.oup.com/jid/article-lookup/doi/10.1086/374419>
- Criscuolo E, Caputo V, Diotti RA, Sautto GA, Kirchenbaum GA, Clementi N. Alternative methods of vaccine delivery: an overview of edible and intradermal vaccines. *J Immunol Res*. 2019;2019. Hindawi Limited. <https://www.hindawi.com/journals/jir/2019/8303648/>
- Zhang GL, Sun J, Chitkushev L, Brusic V. Big data analytics in immunology: a knowledge-based approach. *Biomed Res Int*. 2014; 2014. <https://www.hindawi.com/journals/bmri/2014/437987/>
- Keshavarzi Arshadi A, Webb J, Salem M, Cruz E, Calad-Thomson S, Ghadirian N, et al. Artificial intelligence for COVID-19 drug discovery and vaccine development. *Front Artif Intell*. 2020;3:65. Available from: www.frontiersin.org
- Goodswen SJ, Kennedy PJ, Ellis JT. Vacceed: a high-throughput in silico vaccine candidate discovery pipeline for eukaryotic pathogens based on reverse vaccinology. *Bioinformatics*. 2014;30(16):2381–3. Available from: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu300>
- Tomar N, De RK. Immunoinformatics: an integrated scenario [Internet]. *Immunology*. 2010;131:153–68. Wiley-Blackwell. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2967261/>
- Nascimento AM, Vismari LF, Molina CBST, Cugnasca PS, Camargo JB, De Almeida JR, et al. A systematic literature review about the impact of artificial intelligence on autonomous vehicle safety. *IEEE Trans Intell Transp Syst*. 2020;21:4928–46. Institute of Electrical and Electronics Engineers Inc.

23. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol.* 2017;2(4):230–43. Available from: <http://svn.bmjjournals.org/cgi/doi/10.1136/svn-2017-000101>
24. Heisele B, Ho P, Poggio T. Face recognition with support vector machines: global versus component-based approach. *Proc IEEE Int Conf Comput Vis.* 2001;2:688–94.
25. Arshadi AK, Salem M, Collins J, Yuan JS, Chakrabarti D. Deepmalaria: artificial intelligence driven discovery of potent antimalarials. *Front Pharmacol.* 2020;10. Article number is 1526. <https://www.frontiersin.org/articles/10.3389/fphar.2019.01526/full>
26. Dal Pozzolo A, Boracchi G, Caelen O, Alippi C, Bontempi G. Credit card fraud detection and concept-drift adaptation with delayed supervised information. In: Proceedings of the international joint conference on neural networks. Institute of Electrical and Electronics Engineers Inc.; 2015 international joint conference on Neural networks (IJCNN). IEEE. Killarney, Ireland.
27. Reynisson B, Barra C, Kaabinejadian S, Hildebrand WH, Peters B, Peters B, et al. Improved prediction of MHC II antigen presentation through integration and motif deconvolution of mass spectrometry MHC eluted ligand data. *J Proteome Res.* 2020;19(6):2304–15. Available from: <https://pubmed.ncbi.nlm.nih.gov/32308001/>
28. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* 2020;48(W1):W449–54. Available from: <http://www.cbs.dtu.dk/services/NetMHCIIpan-4.0/>
29. Stafford IS, Kellermann M, Mossotto E, Beattie RM, MacArthur BD, Ennis S. A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases [internet]. *Npj Digit Med.* 2020;3:1–11. <https://doi.org/10.1038/s41746-020-0229-3>. Nature Research.
30. Pulendran B, Li S, Nakaya HI. Systems vaccinology [Internet]. *Immunity.* 2010;33:516–29. Available from: <https://pubmed.ncbi.nlm.nih.gov/21029962/>
31. González-Díaz P, Lee EK, Sorgi S, de Lima DS, Urbanski AH, Silveira EL, et al. Methods for predicting vaccine immunogenicity and reactogenicity. *Hum Vaccin Immunother.* 2020;16(2):269–76. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7062420/>.
32. Chandir S, Siddiqi DA, Hussain OA, Niazi T, Shah MT, Dhama VK, et al. Using predictive analytics to identify children at high risk of defaulting from a routine immunization program: feasibility study. *J Med Internet Res.* 2018;20(9). Available from: <https://pubmed.ncbi.nlm.nih.gov/30181112/>
33. Ahmed SF, Quadeer AA, McKay MR. Preliminary identification of potential vaccine targets for the COVID-19 Coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies. *Viruses.* 2020;12(3). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7150947/>.
34. Chen B, Khodadoust MS, Olsson N, Wagar LE, Fast E, Liu CL, et al. Predicting HLA class II antigen presentation through integrated deep learning. *Nat Biotechnol.* 2019;37(11):1332–43. <https://doi.org/10.1038/s41587-019-0280-2>.
35. Fast E, Altman RB, Chen B. Potential T-cell and B-cell epitopes of 2019-nCoV [Internet]. *bioRxiv;* 2020 [cited 2021 Feb 16]. p. 2020.02.19.955484. <https://doi.org/10.1101/2020.02.19.955484>
36. Yang Z, Bogdan P, Nazarian S. An in silico deep learning approach to multi-epitope vaccine design: a SARS-CoV-2 case study. *Sci Rep.* 2021;11(1):3238. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/33547334>
37. Jespersen MC, Peters B, Nielsen M, Marcatili P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Res.* 2017;45(W1):W24–9. Available from: <https://pubmed.ncbi.nlm.nih.gov/28472356/>
38. Doytchinova IA, Flower DR. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics.* 2007;8(1):4. Available from: <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-4>
39. Gupta S, Kapoor P, Chaudhary K, Gautam A, Kumar R, Raghava GPS. In silico approach for predicting toxicity of peptides and proteins. *PLoS One.* 2013;8(9). Available from: <https://pubmed.ncbi.nlm.nih.gov/24058508/>
40. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. *Nature.* 2020;577(7792):706–10.
41. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. Improved protein structure prediction using predicted interresidue orientations. *Proc Natl Acad Sci U S A.* 2020;117(3):1496–503. Available from: <https://www.pnas.org/content/117/3/1496>
42. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *bioRxiv;* 2020.
43. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. T cells and MHC proteins. 2002 [cited 2021 Feb 16]. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK26926/>
44. Alelign T, Ahmed MM, Bobosha K, Tadesse Y, Howe R, Petros B. Kidney transplantation: the challenge of human leukocyte antigen and its therapeutic strategies [Internet]. *J Immunol Res.* 2018;2018. Hindawi Limited; [cited 2021 Feb 16]. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5859822/>
45. Metzgar RS, Dowell BL, Lachman LB, Jones NH, George FW. Classification of human leukemia by

- membrane antigen analysis with Xenoantisera. *Cancer Res.* 1981;41(11 Part 2):4781–85. <https://pubmed.ncbi.nlm.nih.gov/6794906/>
46. Choo SY. The HLA system: genetics, immunology, clinical testing, and clinical implications. *Yonsei Med J.* 2007;48(1):11–23. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2628004/>.
47. Agarwal RK, Kumari A, Sedai A, Parmar L, Dhanya R, Faulkner L. The case for high resolution extended 6-loci HLA typing for identifying related donors in the Indian subcontinent. *Biol Blood Marrow Transplant.* 2017;23(9):1592–6. Available from: <https://pubmed.ncbi.nlm.nih.gov/28603069/>
48. Tiercy JM. How to select the best available related or unrelated donor of hematopoietic stem cells? [Internet]. *Haematologica.* 2016;101:680–7. Ferrata Storti Foundation. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5013969/>.
49. Stern LJ, Brown JH, Jardetzky TS, Gorga JC, Urban RG, Strominger JL, et al. Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature.* 1994;368(6468):215–21. Available from: <https://pubmed.ncbi.nlm.nih.gov/8145819/>
50. Bhasin M, Raghava GPS. A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes. *J Biosci.* 2007;31–42. Available from: <https://pubmed.ncbi.nlm.nih.gov/17426378/>
51. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, et al. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.* 2015;43(D1): D405–12. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4384014/>.
52. Bassani-Sternberg M, Chong C, Guillaume P, Solleder M, Pak H, Gannon PO, et al. Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allosteric regulating HLA specificity. *PLoS Comput Biol.* 2017;13(8): e1005725. Hertz T, editor. Available from: <https://dx.plos.org/10.1371/journal.pcbi.1005725>
53. Bravi B, Tubiana J, Cocco S, Monasson R, Mora T, Walczak A. Flexible machine learning prediction of antigen presentation for rare and common HLA-I alleles. *bioRxiv* [Internet]; 2020 Apr 25 [cited 2021 Feb 16];2020.04.25.061069. <https://doi.org/10.1101/2020.04.25.061069>
54. Naito T, Suzuki K, Hirata J, Kamatani Y, Matsuda K, Toda T, et al. A multi-task convolutional deep learning method for HLA allelic imputation and its application to trans-ethnic MHC fine-mapping of type 1 diabetes [Internet]. *medRxiv;* 2020 [cited 2021 Feb 16]. p. 2020.08.10.20170522. <https://doi.org/10.1101/2020.08.10.20170522>
55. Wang W, Kiik M, Peek N, Curcin V, Marshall IJ, Rudd AG, et al. A systematic review of machine learning models for predicting outcomes of stroke with structured data. *PLoS One.* 2020;15(6):e0234722. Beiki O, editor. Available from: <https://dx.plos.org/10.1371/journal.pone.0234722>
56. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: a benchmark for molecular machine learning. *Chem Sci.* 2018;9(2):513–30.
57. Walia RR, El-Manzalawy Y, Honavar VG, Dobbs D. Sequence-based prediction of RNA-binding residues in proteins. In: *Methods in molecular biology* [internet]. Humana Press Inc.; 2017. p. 205–35. [cited 2021 Feb 28]. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5796408/>.
58. Bormann N, Koliszak A, Kasper S, Schoen L, Hilpert K, Volkmer R, et al. A short artificial antimicrobial peptide shows potential to prevent or treat bone infections. *Sci Rep.* 2017;7(1):1–14. Available from: <http://aps.unmc.edu/AP/main.php>
59. Salem M, Khormali A, Arshadi AK, Webb J, Yuan J-S. TranScreen: transfer learning on graph-based anti-cancer virtual screening model. *Big Data Cogn Comput.* 2020;4(3):16. Available from: <https://www.mdpi.com/2504-2289/4/3/16>
60. Weimer D, Scholz-Reiter B, Shpitalni M. Design of deep convolutional neural network architectures for automated feature extraction in industrial inspection. *CIRP Ann – Manuf Technol.* 2016;65(1): 417–20.



Artificial Intelligence in Clinical Immunology

101

Aaron Chin and Nicholas L. Rider

Contents

Introduction	1398
Healthcare Data as Big Data	1399
Structuring Healthcare Data	1399
What Is AI and How Does It Work?	1399
AI and the Learning Health System	1400
Applications of AI in Clinical Immunology	1401
Disease Diagnosis: Primary Immunodeficiency	1401
Clinical Decision Support	1402
Candidate Selection for Clinical Trials and Patient Identification Efforts	1402
COVID19 Applications and Impact	1403
Biomarker Identification in Clinical Immunology	1404
Microbiome Analysis	1404
Cytometric Analysis	1405
AI Governance for CI	1405
Ethical Implications in AI	1405
Conclusions	1405
References	1406

Abstract

A. Chin
Department of Medicine, Baylor College of Medicine,
Houston, TX, USA
e-mail: aaron.chin@bcm.edu

N. L. Rider (✉)
Section of Immunology, Allergy & Retrovirology, Texas
Children's Hospital, Baylor College of Medicine, Houston,
TX, USA
e-mail: nlridger@bcm.edu

The field of clinical immunology is expansive and includes both clinical and laboratory science of relevance to human immune health and disease. Across the continuum of host defense, immune homeostasis/regulation, immune genetics, and laboratory immunology we are now seeing the emergence of artificial

intelligence and data science approaches. These computational tools are being leveraged to analyze the inherently large datasets of relevance to clinical immunologists. Here, we outline some recent advances in clinical immunology which have been made possible or are being explored via artificial intelligence. We discuss analysis and use of electronic health record data, human cytometric data, and multiomic data after providing a brief introduction to artificial intelligence in healthcare.

Keywords

Artificial intelligence · Machine learning · Computational biology · Clinical immunology · Data science

Introduction

Since the widespread adoption of electronic health records (EHR) health data are accumulating rapidly [1]. This explosion of EHR data makes health record information truly “big data”; thus, inferences from these datasets present an important opportunity, but require implementation of data science in healthcare and biomedicine to glean insights [2]. In this manner, artificial intelligence (AI)-based computational tools can lead to more robust analysis of clinical and scientific datasets resulting in improvements in patient safety and healthcare quality [2, 3]. For these reasons, AI and machine learning (ML) are finding greater penetration into the healthcare domain space by tackling use cases across multiple spectra [4–6].

While the terms AI and ML are often considered synonyms, AI is more general and refers to computed decision-making and ML is a sub-category of AI [7]. We also make note of the term “augmented intelligence” which is an important distinction in this space, referring to a symbiosis between human efforts and paired digital systems [8]. In ML, a model is programmed to allow machine readable input ingestion and analysis for computer task learning to make predictions with increasing precision over time [2]. In this manner, the machine adjusts weighting of

features to optimize the prediction and minimize error through training and validation.

A useful, recent concept description between humans and computers illuminated a spectrum between fully human and fully computer-driven processes with varying degrees of contributions between the two [9]. This human-computer interaction spectrum incorporates the strata of dataset size as well as complexity of analysis, thereby providing a schema for understanding how humans and machines can interact along the AI continuum. In this schema, one use case may require substantial human domain expertise to build an effective AI, while a deep learning convolutional neural network (CNN) may classify images with minimal human input. These are two extremes of human-computer interaction in AI. Neither is generally superior as it is important to understand that one approach may be more optimally suited to a particular use case. Given the diversity of opportunities and approaches across data science/AI/ML we will use the term “AI” in this chapter as a general reference for such methods across the computational spectrum unless otherwise noted.

Given the aforementioned background, how can and how has AI impacted the field of clinical immunology? Clinical immunology is a field of medicine which concerns itself with the human immune system across host defense, regulation, treatment, and laboratory diagnostics [10]. With this perspective, the impact of AI is noted in clinical immunology by developing methods for improving disease diagnosis via use of computational methods for identifying patients with risk for primary immunodeficiency diseases (PID) [11]. In this way, a risk assessment algorithm can be utilized to scan health plan claims data for the purpose of identifying patients with medical conditions, risks, or trend forecasting to enable early intervention and cost savings. One such example of population-wide risk assessment includes a rule-based analysis of calculated globulin fraction to identify patients who may have hypogammaglobulinemia [12]. Such approaches suggest early penetration of automated computational methods within clinical immunology providing a glimpse into approaches which range from novel disease gene identification to subject selection for

enrollment into clinical trials [13, 14]. In this chapter, we will provide a brief overview of AI followed by an exploration of present clinical immunology approaches. We will cover use cases spanning the advancement of basic biological approaches, improving diagnostic rates and quality of care for patients with immunological disorders.

Healthcare Data as Big Data

Healthcare data are expanding rapidly given the widespread use of EHRs. In every aspect and manner, data generated from healthcare encounters represent “Big Data” (BD). In general, BD implies data which are tremendously voluminous, generated with velocity and highly varied in nature, form, or content [15]. The size and complexity of BD offers challenge and opportunity. These unique features of BD require specific tools to accomplish data inference. These tools, the life cycle of data science, include AI.

Machine learning algorithms are built and trained to ingest diverse data inputs, perform a classification or regression task, and provide output in accordance with that task. The inputs can come from any number of routine clinical, laboratory-, or claims-related data as part of an “example” which is analyzed after labeling them based upon the predefined task [16]. During the model training phase an AI algorithm weights some features more heavily for predicting a specified outcome in accordance with the methodology chosen [2, 7]. Given sufficient performance in validation, an algorithm’s feature weighting may be retained by fixing the algorithm so that new inputs can be classified with the retained weighting [17]. The validated algorithm can then ingest and classify inputs following appropriate structuring.

Structuring Healthcare Data

One way to facilitate the use of AI in healthcare is the construction and use of validated datasets [2]. This requires careful attention to data governance and presents a number of challenges. The

assembly of a useful healthcare dataset first involves the extraction of data elements from the EHR which can have substantial variability, and be unreliable and challenging to extract in systematic fashion [18, 19]. Given the nontrivial task of accessing and making EHR data ready for AI use, clinical teams need to partner with data scientists who can help facilitate information extraction and use by clinicians and laboratory scientists [20]. The expertise of data science teams can be synergistic with clinical and scientific teams to optimize data extraction, creation of validated datasets, and enable rational use and analysis of comprehensive healthcare datasets [21].

What Is AI and How Does It Work?

Machine learning, defined as the process of a computer learning without explicit programming, is the basis of artificial intelligence [2, 22]. In all workflows, AI-algorithms are trained, validated, and then tested prior to their use. There are many machine learning methodologies, each with the goal of automizing computerized learning to complete specific tasks. The common tasks used in AI include classification, regression, and/or clustering. Of these tasks, the most intuitive to understand is “classification” in which the algorithm aims to assign each input into predefined categories (e.g., pneumonia vs. normal study on chest X-ray) [23]. Regression attempts to predict a specific numerical outcome (e.g., length of inpatient stay) and clustering attempts to determine specific categories of output depending on patterns in the data (e.g., finding new classifications for COPD). The inputs used range from structured (e.g., vitals, lab values) to unstructured (e.g., images, text) data [24].

The next step after defining a task is to apply a specific or a set of machine learning methodologies. There are many distinct types of machine learning methods such as deep learning neural networks, random forest, and logistic regression, each with their own advantages. For example, neural networks are a type of deep learning model that mimics the human brain by bidirectionally processing data through a multilayered structure with differentially weighted connections

[25]. This is especially useful in healthcare as additional layers can be added to process large and complex datasets. Once a methodology or combination of methodologies are decided upon, the next step is to “train” the algorithm with a labeled dataset in a process known as supervised learning [24]. If the desired task is to determine the presence of pneumonia on chest X-ray (CXR), the machine would be fed a set CXRs labeled as “normal” or “with pneumonia.” Additional labels (i.e., outlining opacities) may be added to enhance the algorithm’s performance. From there, the algorithm independently learns and discovers specific features to differentiate normal from pathologic X-rays. In contrast to supervised learning, unsupervised learning uses unlabeled data and asks the algorithm to create its own classification scheme. Supervised learning, however, remains the most common form of machine learning [26].

Once the algorithm is trained, the model is put to the test with unseen and unlabeled data. The performance metrics used depends on the type of task desired. For classifier algorithms, commonly used metrics used include recall, precision, positive predictive value, and negative predictive value. Depending on the parameters set on the algorithm, these metrics can be optimized for clinical use. One popular metric, the F1 score, is a composite recall and precision. High F1 scores generally implies good accuracy of an algorithm. Similarly, the area under the receiver-operator curve (RAUC) determines how well an algorithm performs compared to random guessing in a classification task [23]. In the example of the presence of pneumonia on CXR, an algorithm with an AUC value of 1 would differentiate presence versus absence of disease 100% of the time. The other tasks of regression and clustering also have commonly used performance metrics; the discussion of these is outside the scope of this chapter.

AI and the Learning Health System

Ideally a system focused on quality and safety in healthcare will learn and improve over time. This concept was born out of concerns about the overly complex environment in the U.S. rife with

systemic errors, leading to suboptimal outcomes (morbidity and costs). Looking forward with an attempt to address these challenges, the US Institute of Medicine (IOM) considered a different healthcare system which is focused on iterative improvements and optimal outcomes [27]. Circumventing such problems and aiming for quality care produced the notion of the learning healthcare system (LHS) [28]. An LHS was conceived to make it easier to “do the right thing” more often by improving quality and safety, while reducing preventable errors, inefficiency, and sub-optimal processes [29]. In this way, systems could be engineered to redefine healthcare structure and process with intentional use of digital platforms which enable effective clinical decision support systems (CDSS), coordinated care, interoperability, and knowledge acquisition [29].

Visions and early discussions of what is possible brought forth the emergence of concepts from which actual health systems processes were born. These systems were conceived which leverage optimized clinical datasets to drive toward intelligent healthcare systems which are more effective [30–32]. One example of a LHS incorporates time series, structured EHR data for real-time use, and in conjunction with natural language processing (NLP) to forecast clinical events in the intensive care unit (ICU) [33]. Systems of this nature might enable foreshadowing of suboptimal clinical outcomes at the bedside, thereby providing guidance to reduce morbidity or additional patient costs. Another one of many examples includes the use of technology for assessing population risk trends for hospitalization and lockstep adjustment of hospital staffing resources [34]. In this way, population-wide use of wearable sensors can drive real-time health system adjustments while providing precise clinical recommendations for improved health acutely and throughout the lifespan [35]. The intercalation of technology into healthcare processes will move us toward a true LHS by fostering maximum use of data. Such changes have not come easily as they require a culture change and system-wide infrastructure modifications [33].

In order for a holistic HLS to become reality, one that serves patients and clinicians, we need

secure and pervasive data access which take into account the complexity and sensitive nature of healthcare data. The use of common data models and open source tools for clinical data science will then allow for interoperable AI [36, 37]. Along these lines, partitioning a secure development environment containing useful EHR data elements and access to best practice informatic tools will allow for gleaning insights for patients during routine clinical care [38]. Allowing for efficient and secure reuse of clinical data can lead to stepwise improvements in care while extending knowledge via local population health information [39–41]. One important and distinctive feature of health data science and health informatics relates to the highly sensitive nature of the data along with the privatization of platforms (i.e., EHR systems) which capture the bulk of this data. These challenges may be overcome by extracting the most useful data elements and appropriately structuring EHR data within a secure environment where healthcare practitioners and clinical researchers can collaborate together and with data scientists to enable ultimate realization of the HLS.

Applications of AI in Clinical Immunology

Disease Diagnosis: Primary Immunodeficiency

Artificial intelligence has been utilized for improving screening, risk stratification, and disease phenotyping for a variety of conditions [26, 42]. The development of sepsis prediction algorithms to aid clinicians recognize and treat sepsis prior to decompensation is an example of computerized risk stratification in an automatized workflow [43]. In the context of PID, artificial intelligence has been proposed as a diagnostic assist tool for clinicians. The relatively low prevalence and clinical heterogeneity of PID can be a diagnostically challenging, especially for practitioners in nontertiary settings [44]. Work is already underway to leverage large databases to build upon a rule-based algorithm called the

Software for Primary Immunodeficiency Recognition Intervention and Tracking (SPIRIT) to assist in early identification of PID [11]. Using claims data (i.e., ICD codes) the SPIRIT algorithm screens and classifies individuals into low, medium, and high risk for PID categories to help facilitate clinician diagnosis. The addition of supervised learning to this to a population-wide screening tool such as a SPIRIT could aid in earlier recognition of PID, potentially mitigating significant morbidity, mortality, and cost associated with this condition.

In a similar fashion, probabilistic models have been explored as classifiers for various conditions in the field of allergy/immunology [45]. Bayesian networks are an ideal model that combines conditional probabilities and domain expertise to calculate probabilities of interest (e.g., likelihood of disease) [46]. Bayesian networks are built on intuitive nodal structures (e.g., directed acyclic graphs) that represent probabilistic relationships between relevant features and the outcome desired. Using structured EHR data from a large population cohort, a Bayesian network classifying PID versus age-matched controls yielded an AUC of 0.945 [47]. Furthermore, authors demonstrated an 89% accuracy of the model when classifying PID patient into the appropriate International Union of Immunological Societies (IUIS) diagnostic categories. As additional data is incorporated within the algorithm, the Bayesian network can be expanded to include more nodes and connections, thereby improving the model's performance. Incorporation of Bayesian networks and probabilistic models into EHRs will provide clinicians with tools to enable real-time diagnosis [48].

With the advent of electronic health records (EHR) and widespread gene sequencing, clinical phenotyping has become a focus point and a proposed use case for AI algorithms [49]. Over the past decade, distinguishing clinical phenotypes of PID using EHR and genomic data has led to a dramatic increase in the number of distinct PID-associated entities [50]. Machine learning methodologies such as support vector machines (SVM) have been used to discover new PID-associated genes [51]. Integration of

multiomic data, including clinical, genomic, metabolomic, and microbiome data will serve as the basis for molecular phenotyping in foreseeable future [52].

Structured ontologies leveraging structured and unstructured text data have been shown to successfully capture clinical phenotypes through analysis of medical terminology and claims data [53, 54]. For example, ICD-9 codes have been used to capture and characterize PID cases in a statewide population [54]. On the national scale, projects such as the Human Phenotype Ontology (HPO) combines multiple ontologies in the quest to phenotype rare diseases. By linking medical terminology, including terms associated with immunological disorders, HPO attempts to link phenotypes with genomic variants [55, 56]. The use of deep learning methodologies, such as NLP, to extract HPO terms from EHR narratives has been shown to identify causal genes in patients with Mendelian disease [57]. Indeed, similar approaches with NLP have yielded positive results in identifying distinct phenotypes in diseases such as Parkinson's, major depressive disorder, and multiple conditions in the field of allergy/immunology [14, 58, 59]. Ultimately, combining HPO terms, unstructured EHR data mined with NLP, and other structured genomic data to enable artificial intelligence-guided insights and discoveries of immune disorders is the goal of major multi-institutional efforts [60].

Clinical Decision Support

Clinical decision support (CDS) is a subdomain of clinical informatics which focuses on delivering the right information to the right clinician in the most appropriate channel and timing via the EHR [61]. The options for a given CDS intervention range widely, including alerts and custom order sets, among others, including AI-based prediction [62]. Ultimately, CDS is customized according to the given use case and must be formulated with a deep understanding of the particular work flow [63].

Applications of CDS within the domain of CI are beginning to be explored. For example, use of

a rule-based algorithm for assessment of health plan data to provide a risk score for underlying immunological dysfunction [11]. Pilot studies of this nature suggest that automated methods of disease classification could embed within health system workflows to advance clinical analytics processes, as well as improve diagnostic rates and quality of care along the lines of a LHS [26]. Another example includes the prediction of optimal applications for precious clinical resources such as immune globulin (IVIG) in patients with Kawasaki's disease (KD) [64]. In this study, a gradient boosted machine (GBM) was able to predict which KD patients may become resistant to IVIG enabling decision support about best therapeutic options for such patients.

The above studies point toward opportunities to leverage AI to improve care of patients across the spectrum of CI. As of this writing such use cases are only beginning to be explored. As we seek to refine CDS and uncover new applications for using AI to drive insights from the clinical record, CI will continue to benefit from CDS approaches.

Candidate Selection for Clinical Trials and Patient Identification Efforts

Within the field of CI, candidate selection for clinical trials and clinical phenotyping is being improved via AI [65]. For example, NLP can facilitate patient detection for clinical research applications [66]. Patients with lupus and spondylarthritis have been identified with systematic EHR-focused data mining approaches using machine learning [67, 68]. Additionally, use of NLP has been leveraged for understanding atopic dermatitis perception by patients and to better understand the epidemiology of allergic drug reactions [69, 70]. In these two use cases, one can expect to extend the focus of the NLP systems to provide deep phenotyping of the clinical record to identify distinct patient subtypes.

In particular, much has been learned about the electronic phenotype of asthma patients as extracted from clinical records. Patient ascertainment

across the asthma severity spectrum has been accomplished with NLP approaches [14, 71–73]. Asthma control and level of severity can also be mined from clinical note text [74, 75]. These approaches underscore the utility of AI in CI as it pertains to phenotyping patients and identifying subjects for research studies. Additionally, the onerous task of manual patient finding is made more clear from studies that utilize AI to ascertain differences in computable asthma phenotypes from one health system to another [76]. Given a need to assess portability of health records and find patients for a multitude of use cases, AI is being utilized within the domain of CI for these very tasks.

COVID19 Applications and Impact

The greatest public health challenge of our time, the pandemic spread of SARS-CoV-2 and COVID19, is being tackled via AI approaches. First, since the clinical response in COVID19 is heterogeneous, use of artificial neural networks (ANNs) and random forest learning algorithms can be trained to analyze blood count data to diagnose SARS-CoV-2 infection with a high degree of accuracy (AUC 94–95%) [77]. Use of ML to detect COVID19 cases as applied to peripheral blood samples also sheds light on the biology of disease. For example, use of unsupervised ML approaches such as T-REX (Tracking Responders Expanding) can identify rare subpopulations of immune cells in COVID19 patients and provide insights into both early diagnosis as well as the host response to infection.

Understanding how both antibody and cell-mediated immune responses are directed against COVID19 will be important for elucidating biology and driving toward therapies. To better understand humoral immunity against SARS-CoV-2, one group coupled enzyme-linked immunosorbent assay (ELISA) to an ML model for making immunoglobulin binding predictions [78]. Here, the ML model was used to discriminate prior COVID19 infection from binding data. Additionally, use of deep serologic profiling can both

identify exposures as well as find unique viral epitopes for targeting therapies [79]. These applications of AI-based tools can improve therapeutic options for patients and cast light on the biology of this disease.

In addition to diagnosis of individual patients, use of AI and data science may help identify spread and community-level vulnerability [80, 81]. Such approaches may lead to using AI for pattern recognition and augment public health efforts to stem hot spot development. Such tools could also lead to direction of funds in communities where infections are likely to lead to more severe outcomes or enable decision-making about public health worker staffing. Machine learning may also help predict how the SARS-CoV-2 virus spreads and who may be most likely to die from infection [82].

A major focus early and ongoing during the COVID19 pandemic is how best to treat patients. Synthetic drugs and convalescent plasma are important tools in the armamentarium. One example of computed decision-making about COVID19 therapy involves the use of multicriteria decision-making (MCDM) for selecting patients most in need of convalescent plasma (CP) when ill with COVID19 [83]. Also, drug repurposing in the face of a novel pandemic is being approached with AI [84, 85]. From this approach, immune cells can be rapidly profiled for response to viral stress. Additionally, high throughput and AI-based ML can guide drug compound selection and identification based upon their antiviral mechanism of action.

Lastly, vaccine development for COVID19 is being impacted by AI. Use of computational tools and AI for analysis of large datasets is being explored to facilitate the rapid creation of vaccines [86]. Such approaches can leverage the computational efficiency and power of ML to identify novel vaccine targets [87]. In this way, the expanding datasets relating to SARS-CoV-2 genome and proteome can be analyzed. Specifically, coupling reverse vaccinology and ML has allowed for interesting viral-host protein interaction predictions [87]. Ongoing work along these lines will improve our

understanding about the basic biology of current and future pandemics as well as provide prompt options for therapies.

Biomarker Identification in Clinical Immunology

The complex nature of biological systems and their interacting features presents an important opportunity for large dataset analysis with AI. Specifically, identification of biomarkers which yield insights about immune mechanisms and therapies is becoming a subject of great interest and application for ML/AI. Toward understanding mechanisms of allergic skin disease and identification of biosignatures, ML is being used to clarify gene expression patterns from skin biopsies to distinguish allergic contact dermatitis from irritant contact dermatitis [88]. Here, co-expression network analysis and a random forest classifier were used to illuminate biomarker differences between the two forms of immune-mediated skin inflammation. Along these lines, complex diseases such as those which fall within the CI subdomain “atopic diseases” result from polygenic contributors and are amenable to study via computational modeling [89]. The overlap of multiple pathways in atopic dermatitis, for example, has been suggested as a challenging and important use case for advanced analytics approaches [90]. In this way, functional associations and pathway interactions can be drawn with tools such as STRING (<http://string-db.org/>) to tease out mechanistic insights via a multiomics approach.

A multitude of other foci have been described which are amenable to using AI/ML for studying mechanisms. In the field of transplant immunology, ANNs have been created to optimize and personalize immunosuppression through analysis of single-nucleotide polymorphisms (SNPs) [91]. Machine learning has also been used to predict graft versus host disease (GVHD) and response to transplantation [92, 93]. Additionally, a decision tree approach for biomarker ascertainment was created to clarify distinct phenotypes of GVHD and predict outcomes among patients with hematologic malignancy [94].

In addition to host immune approaches, infectious disease-related AI tools are being developed to provide better methods of diagnosis and predictions about outcomes of infectious diseases. Analysis of metabolomics data via a random forest classifier allowed researchers to improve diagnosis of paracoccidiomycosis from other fungal infections [95]. Another study utilized AI/ML in longitudinal fashion to identify unique antigen signatures associated with immunity to malaria [96]. Another example of AI application for biomarker identification of relevance to human disease relates to the prediction of outcomes of *Clostridioides difficile* infection. Use of an elastic net and logistic net facilitated identification of plasma cytokine profiles that predicted clinical endpoints and furthered understanding about immune responses to the bacteria [97].

Microbiome Analysis

Analysis of the microbiome is a separate chapter in and of itself. However, a few mentions about AI biome approaches warrant attention with respect to CI. In particular, machine learning approaches have been applied to predicting outcomes in patients with inflammatory bowel disease (IBD) [98]. Complexity reduction techniques such as unsupervised ML to assess most relevant microbial signatures and classical supervised ML have been used to predict disease severity course and response to treatment, respectively [98, 99]. Specific microbiome signatures have also been analyzed and identified via random forest classifiers to help predict clinical response in Crohn’s disease (CD) with biological agents [100].

Aside from IBD, host pathogen interactions are another use case of relevance in CI. For example a causal mixed graphical model was used to help associate pulmonary microbiome profiles with pneumonia subtypes in ventilated patients [101]. In this way disease-biome correlations can be made via AI to better understand and treat conditions. Similarly, gut microbiome analysis and classification by ML, including other clinical features, was able to predict outcomes in pediatric acute lymphoblastic leukemia (ALL) [102]. Other big data AI approaches toward microbiome study

have provided insights into understanding lung function in cystic fibrosis (CF), methotrexate response in rheumatoid arthritis (RA), delineating asthma endotypes, and helping to clarify the infection-relevant features of immunosenescence in Alzheimer's disease [103–106].

Cytometric Analysis

The human immune system has jurisdiction in all body organs and tissues. It is made up of a variety of cell types which can be studied readily via a sample of blood. To assess cellular constituents of biological samples cytometry utilizes cell biology techniques, physiochemical properties of immune cells as analyzed by lasers (flow cytometry), or mass isotopes (mass cytometry/CyTOF) to perform multiparametric evaluations of cells in solution [107, 108]. In addition to its clinical utility, cytometry approaches have involved AI for more broad applicability. For example, mechanistic immunological domain knowledge and flow/mass cytometry has been fused within an ML platform called the “immunological Elastic-Net” (iEN) [109]. The iEN utilizes prior domain knowledge into Bayesian reasoning for causal modeling which reduces needed cohort training sizes without impairing performance. The end result of this combination of clinical and cytometric data, nestled within a Bayesian framework, is a robust predictive model. Another approach utilized a deep CNN with multiple hidden layers to analyze samples for the purpose of classifying cell populations associated with latent CMV disease [110]. These examples are important applications of AI into the field of CI, specifically immune cytometry, where deep learning can be expertly tailored to recognize cell biomarker populations of relevance to immunological conditions of interest.

AI Governance for CI

Ethical Implications in AI

Notably, data science and AI are extremely useful and will transform healthcare and clinical immunology; however, remembering the distinction

between artificial and augmented intelligence is crucial. Cerrato and Halamka underscore the essence of augmented intelligence when they state “our enthusiastic take on digital innovation should not give readers the impression that AI will ever replace a competent physician. That said, there is little doubt that a competent physician who uses all the tools that AI has to offer will soon *replace* the competent physician who ignores these tools” [62]. Thus it is important that ML model function and prediction is transparent. The “black box” paradigm where mechanisms of output are not clearly evident raises suspicion and leads to mistrust [111]. Past failures are important lessons to heed in making sure that the technology serves patients in a safe and effective manner [112].

Perpetuation of implicit bias must also remain on our minds as we seek to leverage and expand the utility of AI in clinical immunology and healthcare in general [113]. To combat such biases, training datasets must include a diverse population which encompasses geographic disease states, genders, socioeconomic status, and ethnicities that represent the population accurately. Healthcare inequalities remain a challenge to be addressed. These are not isolated to the use of AI, but it is important for us to ensure that new technologies do not worsen discrimination [113, 114]. Yet we already see that healthcare use and outcomes differ across socioeconomic strata, thereby altering the efficacy of CDS systems [115]. For example, extremely well-done and high-profile studies have inadvertently excluded underrepresented minorities in their training algorithms [116]. These examples suggest a field which is maturing and experiencing some growing pains. As such, we need to continue refining the use of AI across healthcare and within the field of clinical immunology.

Conclusions

The field of clinical immunology is growing rapidly with the advent of widespread use of unbiased genomic analysis and more widely available biological therapies [117, 118]. Similarly, use of artificial intelligence in healthcare is expanding to

accommodate large and complex dataset analysis across a variety of use cases [2, 9, 22, 119]. For these reasons, we see a very bright and important future for AI within the domain of CI. Collaboration between data scientists and clinical immunologists will be important. Similarly, training the next generation of clinical immunologists with data science competency is imperative to meld CI domain knowledge and technical skills for solving critical use cases within this specialty.

References

- Obermeyer Z, Lee TH. Lost in thought – the limits of the human mind and the future of medicine. *N Engl J Med.* 2017;377(13):1209–11. <https://doi.org/10.1056/NEJMp1705348>.
- Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med.* 2019;380(14):1347–58. <https://doi.org/10.1056/NEJMra1814259>.
- McGlynn EA, McDonald KM, Cassel CK. Measurement is essential for improving diagnosis and reducing diagnostic error: a report from the Institute of Medicine. *JAMA.* 2015;314(23):2501–2. <https://doi.org/10.1001/jama.2015.13453>.
- Schüssler-Fiorenza Rose SM, et al. A longitudinal big data approach for precision health. *Nat Med.* 2019;25(5):792–804. <https://doi.org/10.1038/s41591-019-0414-6>.
- Norgoet B, Glicksberg BS, Butte AJ. A call for deep-learning healthcare. *Nat Med.* 2019;25(1):14–5. <https://doi.org/10.1038/s41591-018-0320-3>.
- Rajkomar A, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med.* 2018;1: 18. <https://doi.org/10.1038/s41746-018-0029-1>.
- Witten I, Frank E, Hall M, Pal C. Data mining: practical machine learning tools and techniques. Morgan Kaufmann; 2017.
- Autmented intelligence in health care: report 41 of the AMA Board of Trustees. [Online]. https://static1.squarespace.com/static/58d0113a3e00bef537b02b70/t/5b6aed0a758d4610026a719c/1533734156501/AI_2018_Report_AMA.pdf
- Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA.* 2018;319(13):1317–8. <https://doi.org/10.1001/jama.2017.18391>.
- Shearer WT, Fathman CG. 30. Defining the spectrum of clinical immunology. *J Allergy Clin Immunol.* 2003;111(2):S766–73. <https://doi.org/10.1067/mai.2003.88>.
- Rider NL, et al. Calculation of a primary immunodeficiency ‘risk vital sign’ via population-wide analysis of claims data to aid in clinical decision support. *Front Pediatr.* 2019;7:70. <https://doi.org/10.3389/fped.2019.00070>.
- Holding S, Khan S, Sewell WAC, Jolles S, Dore PC. Using calculated globulin fraction to reduce diagnostic delay in primary and secondary hypogammaglobulinaemias: results of a demonstration project. *Ann Clin Biochem.* 2015;52(Pt 3):319–26. <https://doi.org/10.1177/0004563214545791>.
- Sevim Bayrak C, Itan Y. Identifying disease-causing mutations in genomes of single patients by computational approaches. *Hum Genet.* 2020;139(6–7):769–76. <https://doi.org/10.1007/s00439-020-02179-7>.
- Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. *J Allergy Clin Immunol.* 2020;145(2):463–9. <https://doi.org/10.1016/j.jaci.2019.12.897>.
- Chang W, Grady N. NIST big data interoperability framework: volume 1, Definitions, vol. 1. U.S. Dept. of Commerce, National Institute of Standards and Technology; 2019.
- Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, MA: MIT Press; 2016.
- Alpaydin E. Introduction to machine learning. 4th ed. Cambridge, MA: MIT Press; 2020.
- Ford E, Rooney P, Hurley P, Oliver S, Bremner S, Cassell J. Can the use of Bayesian analysis methods correct for incompleteness in electronic health records diagnosis data? Development of a novel method using simulated and real-life clinical data. *Front Public Health.* 2020;8:54. <https://doi.org/10.3389/fpubh.2020.00054>.
- Saria S, Henry KE. Too many definitions of sepsis: can machine learning leverage the electronic health record to increase accuracy and bring consensus? *Crit Care Med.* 2020;48(2):137–41. <https://doi.org/10.1097/CCM.0000000000004144>.
- Dolezel D, McLeod A. Big data analytics in healthcare: investigating the diffusion of innovation. *Perspect Health Inf Manag.* 2019;16(Summer):1a.
- Burke J. Health analytics: gaining insights to transform healthcare. Hoboken: Wiley; 2013.
- Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25(1):44–56. <https://doi.org/10.1038/s41591-018-0300-7>.
- Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC Med Res Methodol.* 2019;19(1):64. <https://doi.org/10.1186/s12874-019-0681-4>.
- Meskó B, Görög M. A short guide for medical professionals in the era of artificial intelligence. *NPJ Digit Med.* 2020;3:126. <https://doi.org/10.1038/s41746-020-00333-z>.
- Esteva A, et al. A guide to deep learning in healthcare. *Nat Med.* 2019;25(1):24–9. <https://doi.org/10.1038/s41591-018-0316-z>.
- Rider NL, Srinivasan R, Khoury P. Artificial intelligence and the hunt for immunological disorders. *Curr Opin Allergy Clin Immunol.* 2020;20(6):565–73. <https://doi.org/10.1097/ACI.0000000000000691>.

27. Kohn L. To err is human: an interview with the Institute of Medicine's Linda Kohn. *Jt Comm J Qual Improv.* 2000;26(4):227–34.
28. Institute of Medicine (US) Roundtable on Evidence-Based Medicine. The learning healthcare system: workshop summary. Washington, DC: National Academies Press (US); 2007.
29. Institute of Medicine (US) and National Academy of Engineering (US) Roundtable on Value & Science-Driven Health Care. Engineering a learning healthcare system: a look at the future: workshop summary. Washington, DC: National Academies Press (US); 2011.
30. Rockowitz S, et al. Children's rare disease cohorts: an integrative research and clinical genomics initiative. *NPJ Genomic Med.* 2020;5:29. <https://doi.org/10.1038/s41525-020-0137-0>.
31. Zhao J, et al. Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction. *Sci Rep.* 2019;9(1):717. <https://doi.org/10.1038/s41598-018-36745-x>.
32. Kuo T-T, Gabriel RA, Cidambi KR, Ohno-Machado L. EXpectation propagation LOgistic REgression on permissioned blockCHAIN (ExplorerChain): decentralized online healthcare/genomics predictive model learning. *J Am Med Inform Assoc JAMIA.* 2020;27(5):747–56. <https://doi.org/10.1093/jamia/oca023>.
33. Marafino BJ, Dudley RA, Shah NH, Chen JH. Accurate and interpretable intensive care risk adjustment for fused clinical data with generalized additive models. *AMIA Jt Summits Transl Sci Proc.* 2018;2017:166–75.
34. Agarwal V, Han L, Madan I, Saluja S, Shidham A, Shah NH. Predicting hospital visits from geo-tagged internet search logs. *AMIA Jt Summits Transl Sci Proc.* 2016;2016:15–24.
35. Agarwal V, Smuck M, Tomkins-Lane C, Shah NH. Inferring physical function from wearable activity monitors: analysis of free-living activity data from patients with knee osteoarthritis. *JMIR MHealth UHealth.* 2018;6(12):e11315. <https://doi.org/10.2196/11315>.
36. Datta S, Posada J, Olson G, Wencheng L, O'Reilly C, Deepa B. A new paradigm for accelerating clinical data science at Stanford. *arXiv.* 2020. [Online]. <https://arxiv.org/abs/2003.10534>
37. Ta CN, Dumontier M, Hripcsak G, Tatonetti NP, Weng C. Columbia Open Health Data, clinical concept prevalence and co-occurrence from electronic health records. *Sci Data.* 2018;5:180273. <https://doi.org/10.1038/sdata.2018.273>.
38. Gombar S, Callahan A, Califf R, Harrington R, Shah NH. It is time to learn from patients like mine. *NPJ Digit Med.* 2019;2:16. <https://doi.org/10.1038/s41746-019-0091-3>.
39. Froelicher D, Misbach M, Troncoso-Pastoriza JR, Raisaro JL, Hubaux J-P. MedCo2: privacy-preserving cohort exploration and analysis. *Stud Health Technol Inform.* 2020;270:317–21. <https://doi.org/10.3233/SHTI200174>.
40. Berliner Senderey A, et al. It's how you say it: Systematic A/B testing of digital messaging cut hospital no-show rates. *PLoS One.* 2020;15(6):e0234817. <https://doi.org/10.1371/journal.pone.0234817>.
41. Schuler A, Callahan A, Jung K, Shah NH. Performing an informatics consult: methods and challenges. *J Am Coll Radiol JACR.* 2018;15(3 Pt B):563–8. <https://doi.org/10.1016/j.jacr.2017.12.023>.
42. Angus DC. Randomized clinical trials of artificial intelligence. *JAMA.* 2020;323(11):1043–5. <https://doi.org/10.1001/jama.2020.1039>.
43. Schinkel M, Paranjape K, Nannan Panday RS, Skyttberg N, Nanayakkara PWB. Clinical applications of artificial intelligence in sepsis: a narrative review. *Comput Biol Med.* 2019;115:103488. <https://doi.org/10.1016/j.combiomed.2019.103488>.
44. Yarmohammadi H, Estrella L, Doucette J, Cunningham-Rundles C. Recognizing primary immune deficiency in clinical practice. *Clin Vaccine Immunol CVI.* 2006;13(3):329–32. <https://doi.org/10.1128/CVI.13.3.329-332.2006>.
45. Ferrante G, Licari A, Fasola S, Marseglia GL, La Grutta S. Artificial intelligence in the diagnosis of pediatric allergic diseases. *Pediatr Allergy Immunol.* 2020;32:405. <https://doi.org/10.1111/pai.13419>.
46. Bayesian artificial intelligence – 2nd Edition – Kevin B. Korb – Ann. <https://www.routledge.com/Bayesian-Artificial-Intelligence/Korb-Nicholson/p/book/9781439815915>. Accessed 06 Feb 2021.
47. Rider NL, et al. PI Prob: a risk prediction and clinical guidance system for evaluating patients with recurrent infections. *PLoS One.* 2021;16:e0237285.
48. McLachlan S, Dube K, Hitman GA, Fenton NE, Kyrimi E. Bayesian networks in healthcare: distribution by medical condition. *Artif Intell Med.* 2020;107:101912. <https://doi.org/10.1016/j.artmed.2020.101912>.
49. Richesson RL, Sun J, Pathak J, Kho AN, Denny JC. Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. *Artif Intell Med.* 2016;71:57–61. <https://doi.org/10.1016/j.artmed.2016.05.005>.
50. Tangye SG, et al. Human inborn errors of immunity: 2019 update on the classification from the International Union of Immunological Societies Expert Committee. *J Clin Immunol.* 2020;40(1):24–64. <https://doi.org/10.1007/s10875-019-00737-x>.
51. Keerthikumar S, et al. Prediction of candidate primary immunodeficiency disease genes using a support vector machine learning approach. *DNA Res.* 2009;16(6):345–51. <https://doi.org/10.1093/dnare/dsp019>.
52. Martorell-Marugán J, et al. Deep learning in omics data analysis and precision medicine. In: Husi H, editor. Computational biology. Brisbane: Codon Publications; 2019.

53. Robinson PN, Haendel MA. Ontologies, knowledge representation, and machine learning for translational research: recent contributions. *Yearb Med Inform.* 2020;29(1):159–62. <https://doi.org/10.1055/s-0040-1701991>.
54. Resnick ES, Bhatt P, Sidi P, Cunningham-Rundles C. Examining the use of ICD-9 diagnosis codes for primary immune deficiency diseases in New York State. *J Clin Immunol.* 2013;33(1):40–8. <https://doi.org/10.1007/s10875-012-9773-1>.
55. Köhler S, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res.* 2019;47(D1):D1018–27. <https://doi.org/10.1093/nar/gky1105>.
56. Köhler S. Improved ontology-based similarity calculations using a study-wise annotation model. *Database.* 2018;2018:bay026. <https://doi.org/10.1093/database/bay026>.
57. Son JH, et al. Deep phenotyping on electronic health records facilitates genetic diagnosis by clinical exomes. *Am J Hum Genet.* 2018;103(1):58–73. <https://doi.org/10.1016/j.ajhg.2018.05.010>.
58. Tracy JM, Özkanca Y, Atkins DC, Hosseini Ghomi R. Investigating voice as a biomarker: deep phenotyping methods for early detection of Parkinson's disease. *J Biomed Inform.* 2020;104:103362. <https://doi.org/10.1016/j.jbi.2019.103362>.
59. Sharma H, et al. Developing a portable natural language processing based phenotyping system. *BMC Med Inform Decis Mak.* 2019;19(Suppl 3):78. <https://doi.org/10.1186/s12911-019-0786-z>.
60. All of Us Research Program Investigators, et al. The 'All of Us' Research Program. *N Engl J Med.* 2019;381(7):668–76. <https://doi.org/10.1056/NEJMsr1809937>.
61. Optimizing strategies for clinical decision support. National Academy of Medicine. <https://nam.edu/optimizing-strategies-clinical-decision-support/>. Accessed 24 Jan 2021.
62. Reinventing clinical decision support: data analytics, artificial intelligence, and diagnostic reasoning. Routledge & CRC Press. <https://www.routledge.com/Reinventing-Clinical-Decison-Support-Data-Analytics-Artificial-Intelligence/Cerrato-Halamka/p/book/9780367186234>. Accessed 24 Jan 2021.
63. Clinical Decision Support | HealthIT.gov. <https://www.healthit.gov/topic/safety/clinical-decision-support>. Accessed 24 Jan 2021.
64. Wang T, Liu G, Lin H. A machine learning approach to predict intravenous immunoglobulin resistance in Kawasaki disease patients: a study based on a South-east China population. *PLoS One.* 2020;15(8):e0237321. <https://doi.org/10.1371/journal.pone.0237321>.
65. Zhang Y, et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). *Nat Protoc.* 2019;14(12):3426–44. <https://doi.org/10.1038/s41596-019-0227-6>.
66. Cai T, et al. EXTraktion of EMR numerical data: an efficient and generalizable tool to EXTEND clinical research. *BMC Med Inform Decis Mak.* 2019;19(1):226. <https://doi.org/10.1186/s12911-019-0970-1>.
67. Jorge A, et al. Identifying lupus patients in electronic health records: development and validation of machine learning algorithms and application of rule-based algorithms. *Semin Arthritis Rheum.* 2019;49(1):84–90. <https://doi.org/10.1016/j.semarthrit.2019.01.002>.
68. Zhao SS, et al. Incorporating natural language processing to improve classification of axial spondyloarthritis using electronic health records. *Rheumatology Oxf Engl.* 2020;59(5):1059–65. <https://doi.org/10.1093/rheumatology/kez375>.
69. Falissard B, et al. Qualitative assessment of adult patients' perception of atopic dermatitis using natural language processing analysis in a cross-sectional study. *Dermatol Ther.* 2020;10(2):297–305. <https://doi.org/10.1007/s13555-020-00356-0>.
70. Banerji A, et al. Natural language processing combined with ICD-9-CM codes as a novel method to study the epidemiology of allergic drug reactions. *J Allergy Clin Immunol Pract.* 2020;8(3):1032–38.e1. <https://doi.org/10.1016/j.jaip.2019.12.007>.
71. Seol HY, et al. Expert artificial intelligence-based natural language processing characterises childhood asthma. *BMJ Open Respir Res.* 2020;7(1):e000524. <https://doi.org/10.1136/bmjresp-2019-000524>.
72. Wi C-I, et al. Natural language processing for asthma ascertainment in different practice settings. *J Allergy Clin Immunol Pract.* 2018;6(1):126–31. <https://doi.org/10.1016/j.jaip.2017.04.041>.
73. Wu ST, Juhn YJ, Sohn S, Liu H. Patient-level temporal aggregation for text-based asthma status ascertainment. *J Am Med Inform Assoc JAMIA.* 2014;21(5):876–84. <https://doi.org/10.1136/amiajnl-2013-002463>.
74. Sohn S, et al. Ascertainment of asthma prognosis using natural language processing from electronic medical records. *J Allergy Clin Immunol.* 2018;141(6):2292–94.e3. <https://doi.org/10.1016/j.jaci.2017.12.1003>.
75. Sauer BC, et al. Performance of a natural language processing (NLP) tool to extract pulmonary function test (PFT) reports from structured and semistructured veteran affairs (VA) data. *EGEMS Wash DC.* 2016;4(1):1217. <https://doi.org/10.13063/2327-9214.1217>.
76. Sohn S, et al. Clinical documentation variations and NLP system portability: a case study in asthma birth cohorts across institutions. *J Am Med Inform Assoc JAMIA.* 2018;25(3):353–9. <https://doi.org/10.1093/jamia/ocx138>.
77. Banerjee A, et al. Use of machine learning and artificial intelligence to predict SARS-CoV-2 infection from full blood counts in a population. *Int Immunopharmacol.* 2020;86:106705. <https://doi.org/10.1016/j.intimp.2020.106705>.

78. Cady NC, et al. Multiplexed detection and quantification of human antibody response to COVID-19 infection using a plasmon enhanced biosensor platform. *Biosens Bioelectron.* 2021;171:112679. <https://doi.org/10.1016/j.bios.2020.112679>.
79. Shrock E, et al. Viral epitope profiling of COVID-19 patients reveals cross-reactivity and correlates of severity. *Science.* 2020;370(6520):eabd4250. <https://doi.org/10.1126/science.abd4250>.
80. Malik YS, et al. How artificial intelligence may help the Covid-19 pandemic: pitfalls and lessons for the future. *Rev Med Virol.* 2020;e2205. <https://doi.org/10.1002/rmv.2205>.
81. Cahill G, Kutac C, Rider NL. Visualizing and assessing US county-level COVID19 vulnerability. *Am J Infect Control.* 2020;49:678. <https://doi.org/10.1016/j.ajic.2020.12.009>.
82. Li M, et al. Identifying novel factors associated with COVID-19 transmission and fatality using the machine learning approach. *Sci Total Environ.* 2020;764:142810. <https://doi.org/10.1016/j.scitotenv.2020.142810>.
83. Albahri OS, et al. Helping doctors hasten COVID-19 treatment: towards a rescue framework for the transfusion of best convalescent plasma to the most critical patients based on biological requirements via ml and novel MCDM methods. *Comput Methods Prog Biomed.* 2020;196:105617. <https://doi.org/10.1016/j.cmpb.2020.105617>.
84. Mirabelli C, et al. Morphological cell profiling of SARS-CoV-2 infection identifies drug repurposing candidates for COVID-19. *BioRxiv Prepr Serv Biol.* 2020. <https://doi.org/10.1101/2020.05.27.117184>.
85. Zhang H, et al. A novel virtual screening procedure identifies pralatrexate as inhibitor of SARS-CoV-2 RdRp and it reduces viral replication in vitro. *PLoS Comput Biol.* 2020;16(12):e1008489. <https://doi.org/10.1371/journal.pcbi.1008489>.
86. Black S, Bloom DE, Kaslow DC, Pecetta S, Rappuoli R. Transforming vaccine development. *Semin Immunol.* 2020;50:101413. <https://doi.org/10.1016/j.smim.2020.101413>.
87. Ong E, Wong MU, Huffman A, He Y. COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning. *Front Immunol.* 2020;11:1581. <https://doi.org/10.3389/fimmu.2020.01581>.
88. Fortino V, et al. Machine-learning-driven biomarker discovery for the discrimination between allergic and irritant contact dermatitis. *Proc Natl Acad Sci U S A.* 2020;117(52):33474–85. <https://doi.org/10.1073/pnas.2009192117>.
89. Lombard C, et al. Clinical parameters vs cytokine profiles as predictive markers of IgE-mediated allergy in young children. *PLoS One.* 2015;10(7):e0132753. <https://doi.org/10.1371/journal.pone.0132753>.
90. Ghosh D, Bernstein JA, Khurana Hershey GK, Rothenberg ME, Mersha TB. Leveraging multilayered ‘omics’ data for atopic dermatitis: a road map to precision medicine. *Front Immunol.* 2018;9:2727. <https://doi.org/10.3389/fimmu.2018.02727>.
91. Fu S, Zarrinpar A. Recent advances in precision medicine for individualized immunosuppression. *Curr Opin Organ Transplant.* 2020;25(4):420–5. <https://doi.org/10.1097/MOT.0000000000000771>.
92. Adom D, Rowan C, Adeniyi T, Yang J, Paczesny S. Biomarkers for allogeneic HCT outcomes. *Front Immunol.* 2020;11:673. <https://doi.org/10.3389/fimmu.2020.00673>.
93. Partanen J, et al. Review of genetic variation as a predictive biomarker for chronic graft-versus-host-disease after allogeneic stem cell transplantation. *Front Immunol.* 2020;11:575492. <https://doi.org/10.3389/fimmu.2020.575492>.
94. Gandelman JS, et al. Machine learning reveals chronic graft-versus-host disease phenotypes and stratifies survival after stem cell transplant for hematologic malignancies. *Haematologica.* 2019;104(1):189–96. <https://doi.org/10.3324/haematol.2018.193441>.
95. de Oliveira Lima E, et al. Metabolomics and machine learning approaches combined in pursuit for more accurate paracoccidioidomycosis diagnoses. *mSystems.* 2020;5(3):e00258-20. <https://doi.org/10.1128/mSystems.00258-20>.
96. Proietti C, et al. Immune signature against *Plasmodium falciparum* antigens predicts clinical immunity in distinct malaria endemic communities. *Mol Cell Proteomics MCP.* 2020;19(1):101–13. <https://doi.org/10.1074/mcp.RA118.001256>.
97. Dieterle MG, et al. Systemic inflammatory mediators are effective biomarkers for predicting adverse outcomes in *Clostridioides difficile* infection. *mBio.* 2020;11(3):e00180-20. <https://doi.org/10.1128/mBio.00180-20>.
98. Tap J, et al. Identification of an intestinal microbiota signature associated with severity of irritable Bowel syndrome. *Gastroenterology.* 2017;152(1):111–23. e8. <https://doi.org/10.1053/j.gastro.2016.09.049>.
99. Douglas GM, et al. Multi-omics differentially classify disease state and treatment outcome in pediatric Crohn’s disease. *Microbiome.* 2018;6(1):13. <https://doi.org/10.1186/s40168-018-0398-3>.
100. Doherty MK, et al. Fecal microbiota signatures are associated with response to Ustekinumab therapy among Crohn’s disease patients. *mBio.* 2018;9(2). <https://doi.org/10.1128/mBio.02120-17>.
101. Kitsios GD, et al. Respiratory microbiome profiling for etiologic diagnosis of pneumonia in mechanically ventilated patients. *Front Microbiol.* 2018;9:1413. <https://doi.org/10.3389/fmcb.2018.01413>.
102. Nearing JT, et al. Infectious complications are associated with alterations in the gut microbiome in pediatric patients with acute lymphoblastic leukemia. *Front Cell Infect Microbiol.* 2019;9:28. <https://doi.org/10.3389/fcimb.2019.00028>.
103. Zhao CY, et al. Microbiome data enhances predictive models of lung function in people with cystic fibrosis.

- J Infect Dis. 2020. <https://doi.org/10.1093/infdis/jiaa655>.
104. Artacho A, et al. The pre-treatment gut microbiome is associated with lack of response to methotrexate in new onset rheumatoid arthritis. *Arthritis Rheumatol*. Hoboken NJ. 2020. <https://doi.org/10.1002/art.41622>.
 105. Lejeune S, et al. Childhood asthma heterogeneity at the era of precision medicine: modulating the immune response or the microbiota for the management of asthma attack. *Biochem Pharmacol*. 2020;179:114046. <https://doi.org/10.1016/j.bcp.2020.114046>.
 106. Haran JP, et al. Alzheimer's disease microbiome is associated with dysregulation of the anti-inflammatory P-glycoprotein pathway. *mBio*. 2019;10(3). <https://doi.org/10.1128/mBio.00632-19>.
 107. McKinnon KM. Flow cytometry: an overview. *Curr Protoc Immunol*. 2018;120:5.1.1–11. <https://doi.org/10.1002/cpim.40>.
 108. Kay AW, Strauss-Albee DM, Blish CA. Application of mass cytometry (CyTOF) for functional and phenotypic analysis of natural killer cells. *Methods Mol Biol* Clifton NJ. 2016;1441:13–26. https://doi.org/10.1007/978-1-4939-3684-7_2.
 109. Culos A, et al. Integration of mechanistic immunological knowledge into a machine learning pipeline improves predictions. *Nat Mach Intell*. 2020;2(10):619–28. <https://doi.org/10.1038/s42256-020-00232-8>.
 110. Hu Z, Tang A, Singh J, Bhattacharya S, Butte AJ. A robust and interpretable end-to-end deep learning model for cytometry data. *Proc Natl Acad Sci U S A*. 2020;117(35):21373–80. <https://doi.org/10.1073/pnas.2003026117>.
 111. Castelvecchi D. Can we open the black box of AI? *Nature*. 2016;538(7623):20–3. <https://doi.org/10.1038/538020a>.
 112. IBM's Watson recommended 'unsafe and incorrect' cancer treatments, STAT report finds. <https://www.beckershospitalreview.com/artificial-intelligence/ibm-s-watson-recommended-unsafe-and-incorrect-cancer-treatments-stat-report-finds.html>. Accessed 24 Jan 2021.
 113. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med*. 2018;178(11):1544–7. <https://doi.org/10.1001/jamainternmed.2018.3763>.
 114. Stringhini S, et al. Socioeconomic status and the 25 × 25 risk factors as determinants of premature mortality: a multicohort study and meta-analysis of 1·7 million men and women. *Lancet Lond Engl*. 2017;389(10075):1229–37. [https://doi.org/10.1016/S0140-6736\(16\)32380-7](https://doi.org/10.1016/S0140-6736(16)32380-7).
 115. Arpey NC, Gaglioti AH, Rosenbaum ME. How socio-economic status affects patient perceptions of health care: a qualitative study. *J Prim Care Community Health*. 2017;8(3):169–75. <https://doi.org/10.1177/2150131917697439>.
 116. Esteva A, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115–8. <https://doi.org/10.1038/nature21056>.
 117. Chinn IK, Orange JS. A 2020 update on the use of genetic testing for patients with primary immunodeficiency. *Expert Rev Clin Immunol*. 2020;16(9):897–909. <https://doi.org/10.1080/1744666X.2020.1814145>.
 118. Stray-Pedersen A, et al. Primary immunodeficiency diseases: genomic approaches delineate heterogeneous Mendelian disorders. *J Allergy Clin Immunol*. 2017;139(1):232–45. <https://doi.org/10.1016/j.jaci.2016.05.042>.
 119. Norgeot B, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med*. 2020;26(9):1320–4. <https://doi.org/10.1038/s41591-020-1041-y>.



Lukas Wisgrill, Paulina Werner, Vittorio Fortino, and
Nanna Fyhrquist

Contents

Introduction	1412
General Principles of Machine Learning	1413
The Allergic March: Can We Predict the Direction?	1414
The Needle in the Hay: ML-Approaches for Allergy Biomarker Discovery	1415
Finding a Pattern: Disease Subtyping for Clinical Management	1416
Bridging the Gap: Insights into Disease Biology by Integrative Analysis	1418
Challenges and Promising Research Directions in ML Applications	1418
References	1420

Abstract

Allergic disorders are highly complex and heterogeneous in the sense of pathogenesis, molecular mechanisms, diagnostic tools, and

treatment success. The advent of novel systems immunology approaches unraveled new dimensions in the field of precision allergology to improve patient well-being and outcome. However, a major challenge within the field is to deconvolute the vast and diverse datasets to extract innovative and valuable information out of clinically heterogeneous patient populations. Machine-learning algorithms offer new tools to gain insights into high-dimensional omics datasets as well as clinical patient records, providing data-driven methodologies to decipher hidden features and patterns behind big data. These methods have a high translational aspect since newly identified mechanisms or endotypes can lead to innovative treatment approaches. This chapter gives a short overview of the current literature as well as an overview of the used machine-learning

L. Wisgrill
Division of Neonatology, Pediatric Intensive Care and
Neuropediatrics, Comprehensive Center for Pediatrics,
Department of Pediatrics and Adolescent Medicine,
Medical University of Vienna, Vienna, Austria
e-mail: lukas.wisgrill@meduniwien.ac.at

P. Werner · N. Fyhrquist (✉)
Institute of Environmental Medicine, Karolinska Institutet,
Stockholm, Sweden
e-mail: paulina.werner@ki.se; nanna.fyhrquist@ki.se

V. Fortino
Institute of Biomedicine, University of Eastern Finland,
Kuopio, Finland
e-mail: vittorio.fortino@uef.fi

approaches in modern allergology. Furthermore, representative examples are given to illustrate the rationale behind the different machine-learning algorithms used in the research field of allergology.

Keywords

Machine learning · Disease risk prediction · Biomarker discovery · Disease endotypes · Allergy diagnostics · Disease prognosis · Precision allergology

Introduction

Allergic disorders, which entail a number of hypersensitivity conditions with exaggerated immune responses toward environmental exposures, are a major health problem particularly during childhood, showing an increasing prevalence worldwide [1]. The growing incidence in allergic diseases is associated with changes in lifestyle and infection control, often referred to as the “hygiene hypothesis” [2], later extended by the “Old friends” [3] and “Biodiversity” [4] hypotheses. In addition, the continuous introduction of new potentially sensitizing substances in various consumer products, adds to the already extensive repertoire of sensitizers in our living environment, leading to increased risks of sensitization [5].

Allergic disorders that result from exaggerated IgE-mediated immune responses toward harmless environmental substances are known as atopic disorders, manifesting in diseases such as atopic dermatitis (AD), food allergies (FA), allergic rhinitis (AR), and asthma [1]. This disease sequence is also referred to as the “allergic march.” Contrarily, other allergic disease entities such as allergic contact dermatitis (ACD) is IgE-independent, exhibiting strong T cell-mediated reactions from direct topical exposure to environmental substances such as metals or rubbers [6]. Although ACD and atopic diseases display different molecular mechanisms, patients with AD often suffer from ACD as well, making clinical diagnosis and treatment more complex [7].

AD is the most common form of eczema in childhood, manifesting as a chronic, relapsing itchy rash. Constant itching and stigmata associated with visible skin disease can have a major impact on the quality of life of individuals suffering from AD [8]. FA, which is an abnormal immune response to food, has a growing prevalence worldwide. The symptoms range from mild to severe, including itchiness, swelling of the tongue, vomiting, diarrhea, trouble breathing, and low blood pressure. Disease management heavily relies on avoidance and emergency preparedness [9]. Asthma is a common long-term inflammatory disease of the airways, characterized by airflow obstruction and easily triggered bronchospasms. Symptoms of asthma include episodes of wheezing, coughing, and shortness of breath [10], whereas AR, which is an inflammation of the nose, includes symptoms such as a runny or stuffy nose, sneezing, and red, itchy, watery eyes. In subjects with asthma or AR, or both, quality of life is greatly impaired by physical limitations, difficulties with daily activities and poorer mental well-being [11].

In recent years, extensive progress has been made in identifying pathomechanisms of allergic disease, refining the definition of disease phenotypes, identifying novel biomarkers for a better perspective of the patients’ pathology, and for the development of preventive strategies as well as new treatment modalities [12]. However, many important questions remain unsolved regarding disease mechanisms, disease susceptibility, personalized patient treatment, treatment efficacy, and treatment-induced adverse side effects and long-term effects.

The widespread availability of new high-throughput technologies, such as genomics, proteomics, transcriptomics, and lipidomics, has greatly increased the amount of data available for the discovery of diagnostic or prognostic markers and therapeutic targets in allergic diseases [13]. However, using conventional techniques, it is a challenge to combine the high number and mixtures of data sources to find meaningful patterns. Machine learning (ML) algorithms offer an opportunity to more easily analyze all the available data, significantly speeding up, e.g., the discovery of drug

candidates or biomarkers for diagnosis. Moreover, ML can automate the complicated statistical work needed to predict a patients' response to a particular treatment. Therefore, ML methods are expected to significantly empower management of allergic disease and related research in the near future.

General Principles of Machine Learning

ML is a branch of Artificial Intelligence (AI) focusing on developing algorithms that can learn from data, without being explicitly programmed. The process by which ML algorithms learn from the data can be classified in four major categories: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning.

In supervised learning, the ML algorithm maps an input to an output based on a known set of sample data with known characteristics (e.g., clinical data, biomarkers, etc.) and labels (e.g., patients with a skin disease and healthy controls). A typical supervised learning task is classification, which can be particularly useful for diseases with known subtypes that require proper subtype classification for adequate therapeutic measures. For instance, diagnostic discrimination between the allergic and irritant form of contact dermatitis (CD) is of importance for the clinical disease management, and supervised ML methods have been employed to identify diagnostic biomarkers to distinguish these two types of CD [14].

Similarly to supervised learning, the goal of unsupervised learning is to place input data into categories. The main difference is that the input data are not labeled in unsupervised learning, and therefore, the ML algorithms try to categorize data (e.g., patient samples) based on their inherent features (e.g., gene expression signatures, clinical features, etc.). Unsupervised techniques are instrumental in finding hidden patterns or grouping the data based on the identified patterns. Typically, unsupervised learning tasks include a) dimensionality reduction, which reduces the initial number of features (e.g., number of genes quantified by using microarrays or next-

generation sequencing (NGS) technologies); and b) clustering, which aims to organize different variables or samples into groups based on observable similarities. For example, patients exhibiting similar clinical or molecular features could provide precise identification of patient subsets for tailored therapeutics [15]. Asthma, for instance, is a heterogeneous disease comprising a number of subtypes, which may share similar molecular patterns (sometimes referred to as endotypes) or similar clinical characteristics (phenotypes). The use of unsupervised clustering is a very common practice in identifying subtypes of asthma based on observable molecular and clinical characteristics [16].

Semi-supervised machine learning is a hybrid strategy of ML that combines aspects of supervised and unsupervised learning. In this strategy, large and unlabeled data sets are combined with a small number of labeled inputs in order to alleviate the problem of data labeling. For example, patients in certain types of patient data (e.g., electronic health records – EHR) may be largely unlabeled as it can be expensive and time consuming for clinicians to label the data at research quality. In addition, this task can be tedious and can lead to labeling errors. For that reason, semi-supervised learning techniques are often preferred in order to prevent or, at least, reduce the risk of human errors in sample labeling. Moreover, deep learning-based approaches (e.g., denoising auto-encoders) have been successfully applied for the training of semi-supervised classifiers that enable the use of large un-annotated EHR-data in order to improve classification accuracy [17].

Last, reinforcement learning is a distinct category of ML alongside supervised and unsupervised learning, and their hybrid, semi-supervised learning. In contrast to supervised learning, reinforcement learning does not depend on labeled input or output, or the revision of near-optimal decisions. Instead, the mission is to find a balance between searching the unknown and processing current knowledge. Reinforcement algorithms use dynamic programming techniques, and target large decision processes, where exact mathematical models become infeasible. In the medical field, decision analysis and techniques of reinforcement

learning could theoretically mitigate challenges such as harm caused to patients due to cognitive and judgment errors made under time constraints, the uncertainty regarding the patient's diagnosis, or as a result of doubts regarding the predicted response to treatment. In the medical context, the available methods are poorly understood and rarely used clinically [18]. Therefore, promoting a better understanding of these approaches could substantially support clinical decision making, facilitating, for instance, the prediction of asthma exacerbations [19].

The Allergic March: Can We Predict the Direction?

As mentioned above, allergic diseases arise from an inappropriate immune response against environmental antigens. Although allergic diseases can seemingly affect specific organs, the different clinically observed entities are also systemic with partial overlapping etiologies. Allergic children have an estimated ~10-fold higher risk to develop further allergic conditions in later life [20]. These stepwise developing allergic co-occurrences are currently referred to as the "allergic march." The "classic" allergic march starts with AD in infancy, followed by the development of FA and subsequently AR and asthma [21]. Furthermore, patients with such progressive allergic history are at increased risk of developing eosinophilic esophagitis [22]. However, the individual clinical course can vary between patients, underlining the importance of novel tools to predict and characterize allergic diseases. In the advent of systems-level omics techniques, dissecting the clinical phenotype into allergic endotypes may be a promising approach to improve our understanding of allergic diseases in the context of precision medicine [23].

Various single- and cross-omics approaches have been used to decipher the complex and interwoven biological data layers of AD [24]. This inflammatory skin disease is primarily characterized by epidermal barrier dysfunction and altered immune responses, clinically presenting as itchy and eczematous skin lesions. Deeper insights into

the pathophysiology of AD might unravel novel treatment targets for AD patients as well as decipher unknown mechanisms of the allergic march. For example, AD cross-omics overlap signatures at the tissue level identified enrichment for skin and esophagus, highlighting a potential link between AD and the progression toward FA [24]. Recent studies further suggest that skin and gut are linked via the modulation of the immune environment by the human microbiome, and that certain microbiome patterns can determine sensitization to food and environmental allergens, and may predispose the host to develop an allergy [25–27]. Therefore, cross-omics approaches combined with ML tools have the power to provide novel disease models as well as predicting disease risk.

Example 1 Disease Risk Prediction with ML-Driven Models

Important tasks of precision medicine include the classification of patients according to disease risk as well as the estimation of disease probability, which is often accomplished by the use of different univariate statistical methods such as Student's t-test or Fisher's exact test. However, ML-based approaches could offer some advantages over classical statistical methods as it allows statistical testing in a multivariate fashion. For instance, when considering genome-wide association studies (GWAS), the goal is to identify a subset of genes and causal variants that determine the susceptibility to disease. The basic association test to discover single genetic variants (and genes) that are informative for assessing disease risk is based on comparing allele frequencies of single nucleotide polymorphisms (SNPs) between cases and controls. Strongly associated genetic variants are then combined into a Polygenic Risk Score (PRS), which is defined as a linear combination of individual SNP effects determined by using linear or logistic regression [28]. Although the PRS depends on statistical techniques that are commonly used in ML, it is built on the assumption that the underlying data are normally distributed, the data observations are non-correlated, and that they include additive and independent predictor effects [29]. However, PRSs rely on statistical

association tests that tend to select a very small set of SNPs with large effect sizes [30]. This means that the models tend to oversimplify the biological processes underlying complex diseases by ignoring SNPs that make a smaller contribution to the phenotype [30, 31]. Moreover, in the search for associated sets of SNPs, many causal variants may be missed [32]. More advanced ML approaches, such as Support Vector Machine (SVM) or Random Forest (RF), can generate more accurate predictions, by computing the complex relationships between risk SNPs (either a large subset of SNPs or all SNPs) and associating them with complex disease phenotypes, such as asthma. For example, recent studies have shown that the use of SVM- and RF-based classifiers can improve SNP-based predictive models of asthma [33, 34]. The predictive performances of ML-driven models are evaluated by using repeated, cross-fold validations and evaluation metrics such as the AUC, the positive and negative predictive values, or the coefficient of determination R². The main advantages of the ML-driven approaches are that they account for inter-SNP correlations, rather than assuming independence across the individual SNPs (Fig. 1).

The Needle in the Hay: ML-Approaches for Allergy Biomarker Discovery

The accurate diagnosis of allergies requires validation of allergen sensitization as well as detailed description of antigen exposure to the putative allergen. Skin tests, especially the Skin Prick Test, represent the current gold standard for the diagnosis and management of IgE-mediated allergic diseases. However, in some cases (young children below 2 years of age, patients with an unstable medical condition or at high risk of anaphylaxis, etc.) blood testing, such as in vitro serum IgE (sIgE) detection, might be preferred. Nevertheless, those blood tests need to be standardized and accurate, and especially FA panels need to be approved for cost-effectiveness and specificity [35].

FA encompasses a range of different food hypersensitivities as well as clinical phenotypes, including mild to severe allergic reactions.

Currently, it is estimated that nearly 8% of children and 3–5% of adults are affected by FAs and the incidence is increasing worldwide [36, 37]. Interestingly, the majority of children with a positive Skin Prick Test or sIgE results do not suffer from FA. Therefore, oral food challenges are the current gold standard to diagnose clinical reactivity against food allergens. However, this procedure is expensive, stressful for the child, time-consuming, and potentially riskful for the patient by developing anaphylaxis [38]. Consequently, a suitable blood test with high accuracy and minimal risk (e.g., blood drawing) could solve current diagnostic problems in FA. In recent years, epigenetic markers have been discovered to diagnose FA, showing a high accuracy when ML-driven approaches are used to identify biomarkers [39–41].

Example 2 Discovery of Biomarkers in Allergic Diseases Using Feature Selection

The implementation of precision medicine in managing chronic diseases strongly relies on the definition of biomarkers. The use of biomarkers in medicine is based on the presumption that there are measurable markers that can be used as a proxy of a biological process by, for instance, being an indicator of the presence of disease [42]. Even though they are often referred to as diagnostic tools, biomarkers can serve a wide range of clinical applications including population screening, prognosis, monitoring, and prediction of therapeutic response or toxicity [43]. In modern translational research, the development of novel biomarkers heavily relies on use of omics technologies as an important data source for the discovery of new diagnostic and prognostic biomarkers. These technologies allow precise measurements for millions of within-cell molecular features, such as genes, mRNA, proteins, and metabolites in a specific biological sample, which results in a plentitude of molecules that may be selected to composite biomarker panels for prediction tasks in precision medicine. For this reason, biomarker discovery from omics data is a laborious task. Moreover, it is important that this process leads to the selection of a subset of molecular markers that identify target classes of clinical

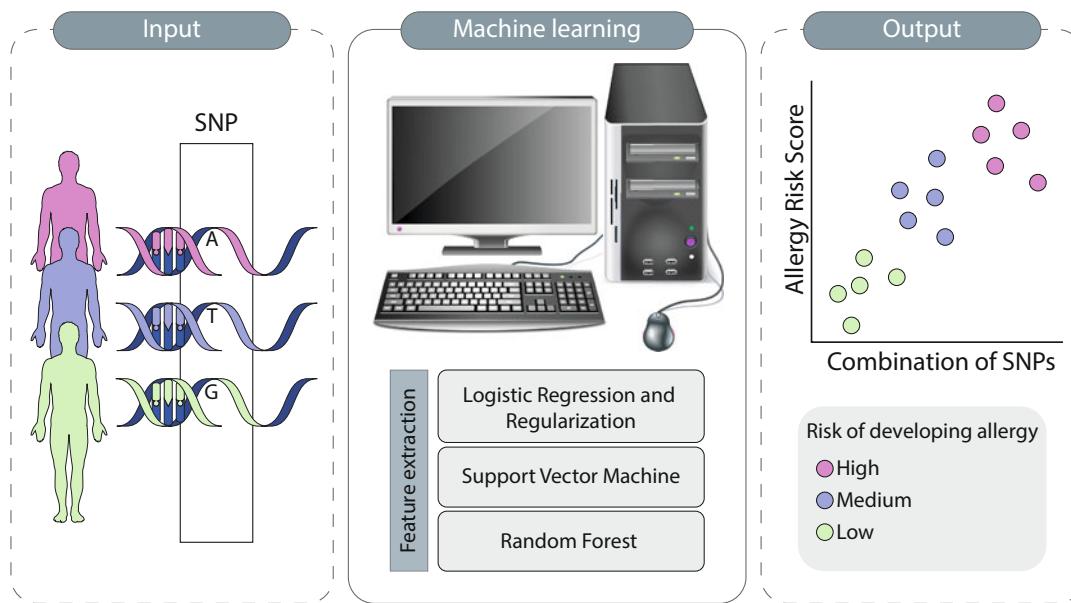


Fig. 1 Example of ML methods that can be applied for disease risk prediction in allergology. Data obtained from genome wide association studies (GWAS) such as allele frequencies from cases and controls (leftmost column) are combined and the contributing features, such as disease-associated SNPs, can be extracted through ML-based

approaches (middle column). As a result, the risk of allergy based on the combination of disease-associated SNPs can then be predicted as illustrated in the plot in the rightmost column. (This figure was created using vectors from Servier Medical Art, under a Creative Commons Attribution 3.0 Unported License; <http://smart.servier.com>)

cases [44]. Indeed, the discovery of biomarkers from omics is typically modeled as a feature selection problem, where the aim is to identify the most discriminating features for a given classification. For example, when identifying biomarkers to distinguish between the allergic and irritant CD or classifying skin diseases, ranging from melanoma to psoriasis, a feature selection methodology may be applied. In ML, this is the process of selecting a subset of relevant features (variables, predictors) to be used in the model construction. Common approaches for feature selection are filter, wrapper, and the embedded methods. Filter methods measure the relevance of features (or the potential biomarkers) by assessing their correlation with the target variable and wrapper methods quantify the “usefulness” of a set of features by evaluating a machine learning model on it. Finally, the embedded methods are based on ML algorithms having a feature selection process that is embedded within the model fitting process.

A classic wrapper method is based on the use of Genetic Algorithms (GAs). An example of

GA-based wrapper feature selection methods is GARBO [45], which is an omics-driven approach that is able to optimize accuracy and set size simultaneously for selection of reliable biomarkers. GARBO has been recently used to identify robust gene expression-based biomarkers for the discrimination between allergic and irritant CD [14] (Fig. 2).

Finding a Pattern: Disease Subtyping for Clinical Management

An important feature of precision medicine is disease subtyping, which can lead to a better understanding on how to optimally treat individual patients. Disease subtyping aims to identify subpopulations of patients showing similar phenotypic, clinical or molecular characteristics. When the subtypes can be linked to distinct pathophysiological mechanisms of the disease, they are also called endotypes. The endotype may also include data from environmental exposures, and

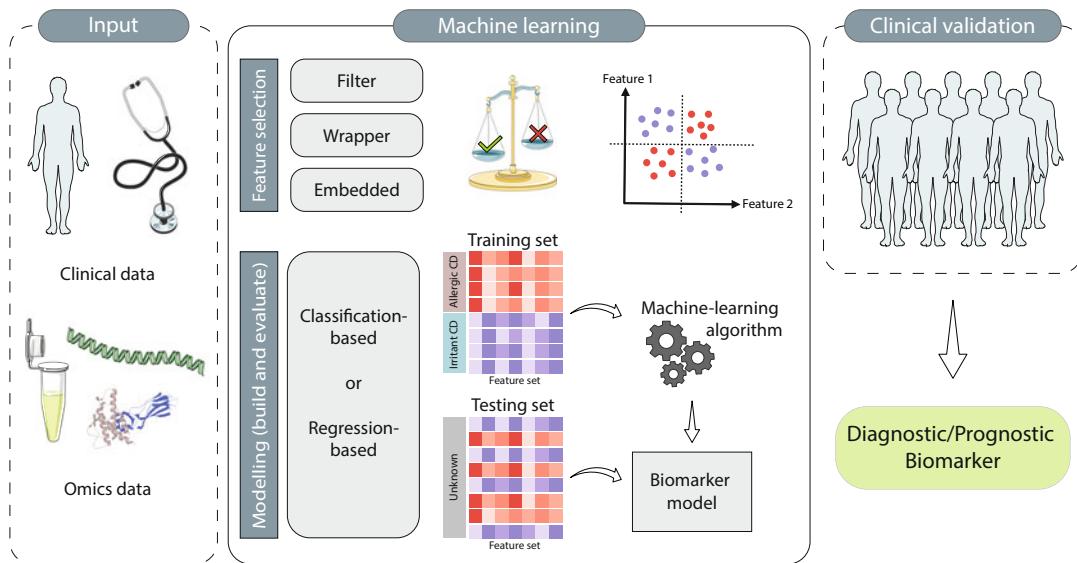


Fig. 2 Example of how feature selection and classification analysis can be applied for biomarker discovery from omics data. The classical steps of ML consist of feature extraction, model learning, and model evaluation. First, clinical and omics data are collected (leftmost column) and features that can classify clinical conditions are extracted by feature selection methods (middle column). Next, models of extracted features are built using a training set

features of the human microbiome. Together with the clinical phenotype, the endotype is now considered a fundamental tool for stratification of these highly complex diseases.

AD is still largely considered and managed as a single disease, lacking tailored prevention and therapeutic strategies. However, the clinical phenotype of AD is highly heterogeneous, and will benefit from dissection into more homogeneous subgroups for refined management [46, 47]. A prerequisite for precision medicine in AD is the discovery and validation of reliable biomarkers, and ML methodologies will help to merge the increasing amounts of data to define reliable endotypes. One important, unanswered question is whether there are endotypes in AD related to skin microbiome diversity or specific bacterial strains. A recent study used supervised learning based on Random Forest classification to pinpoint AD relevant skin communities of microorganisms, discovering two subtypes of AD, characterized by either high or low abundance of *Staphylococcus aureus*, with distinct consequences for disease severity and expression of

and the prediction accuracy is tested using a testing set of unknown samples. Finally, the biomarker model is validated clinically to ensure its generalizability and clinical validity (rightmost column). (This figure was created using vectors from Servier Medical Art, under a Creative Commons Attribution 3.0 Unported License; <http://smart.servier.com>)

inflammatory and skin barrier function related host genes [48].

In contrast to AD, the model of a single entity in asthma is no longer in use; Asthma is currently considered as an umbrella diagnosis for several diseases, driven by distinct mechanisms (endotypes) and with variable clinical manifestations (phenotypes). Asthma is broadly divided into type 2 (T2)-high or T2-low endotypes, depending on the extent of the involvement of Th2 type immune signaling. In the management of asthma, proper definition of endotypes and their sub-endotypes is essential due to inherent therapeutic and prognostic implications [49].

Example 3 Disease Phenotyping and Endotyping by Unsupervised Cluster Analysis

As mentioned in section “[The Allergic March: Can We Predict the Direction?](#),” distinct subtypes of asthma can be discovered by applying unsupervised clustering algorithms [50–52]. Traditionally, diseases have been stratified into disease subtypes based on their clinical manifestations, or the

so-called phenotypes. For example, distinct clinical subtypes of asthma have been identified using unsupervised hierarchical cluster analysis or latent class analysis of cohorts based on phenotypic data, including questionnaire data, demographic data (sex, race, age), disease severity (age of onset, asthma duration), and physiologic measures (lung function, atopy) [51, 52].

Although the clustering of patients based on phenotypic data may be relevant, the resulting subgroups often show inconsistencies, which can be due to differences in the demographic or clinical characteristics of the populations studied [53]. Moreover, a single phenotype may consist of multiple molecular endotypes [13].

Molecular endotypes can be characterized by using -omics approaches, including transcriptomics, epigenomics, microbiomics, metabolomics, and proteomics. Single- and multi-omics profiles of patients with asthma can help identify endotypes, or subtypes of disease based on distinct pathophysiological mechanisms [54, 55]. For instance, the k-means clustering algorithm was applied to the gene expression profiles of PBMC from asthmatic children in order to define novel endotypes from asthmatic children [56]. Metabolomic and epigenetic data have also been used, in combination with clustering algorithms, for the identification of novel asthma endotypes [57, 58] (Fig. 3).

Bridging the Gap: Insights into Disease Biology by Integrative Analysis

Most studies using omics data to identify biological targets and pathways involved in asthma and allergic disease pathogenesis have primarily focused on single-omics subtypes. Although important insights have been gained from these studies, they provide only a partial view of the disease process. An integrative approach, where features from multiple omics data layers are modeled as a set, offers an opportunity to study complex biological processes holistically [54]. The recent arrival of new high-throughput techniques, a growing availability of data from large patient cohorts, and the development of several promising methods for data integration, pave

the way for improved mechanistic insight, disease subtyping, and biomarker prediction. For instance, a study examining 361 children of 3 years of age investigated the effects of the exposome on the risk of developing childhood asthma through coinertia and sparse correlation analysis of intestinal metabolome, intestinal microbiota, plasma metabolome, and diet [59]. In this study, several intestinal metabolites, specific bacterial taxa (such as the family Christensenellaceae) as well as plasma metabolites and a meat-rich diet were found to be associated with the development of childhood asthma. In another study, metabolomic data from plasma samples, peripheral blood CD4+ methylation and transcriptomic profiles, and genome-wide genotypes from 20 asthmatic patients were integrated to create a conditional Gaussian Bayesian network [60]. As a result, the integrative approach revealed that two pathways, specifically arachidonic acid metabolism and linoleic acid metabolism, may be of importance for asthma control. Furthermore, the analysis implicated altered sphingolipid metabolism as a potential underlying feature of uncontrolled asthma. Both of these examples illustrate the potential of successful generation of meaningful insights into pathophysiological mechanisms of asthma by integration of multiple omics dimensions to create a more complete picture of complex biological systems.

Challenges and Promising Research Directions in ML Applications

With allergies representing exceedingly heterogeneous disorders, it is of utmost importance to decipher key biological pathways to improve the current knowledge of diagnosis and treatment of such diseases. However, ML approaches are challenging, not the least from a technical point of view. One important issue is the risk of overfitting due to small sample sets, which include a high number of features – a common trait of biomedical data (e.g., EHR, omics, imaging data, etc.). The problem of overfitting arises when ML generates a predictive model that performs perfectly

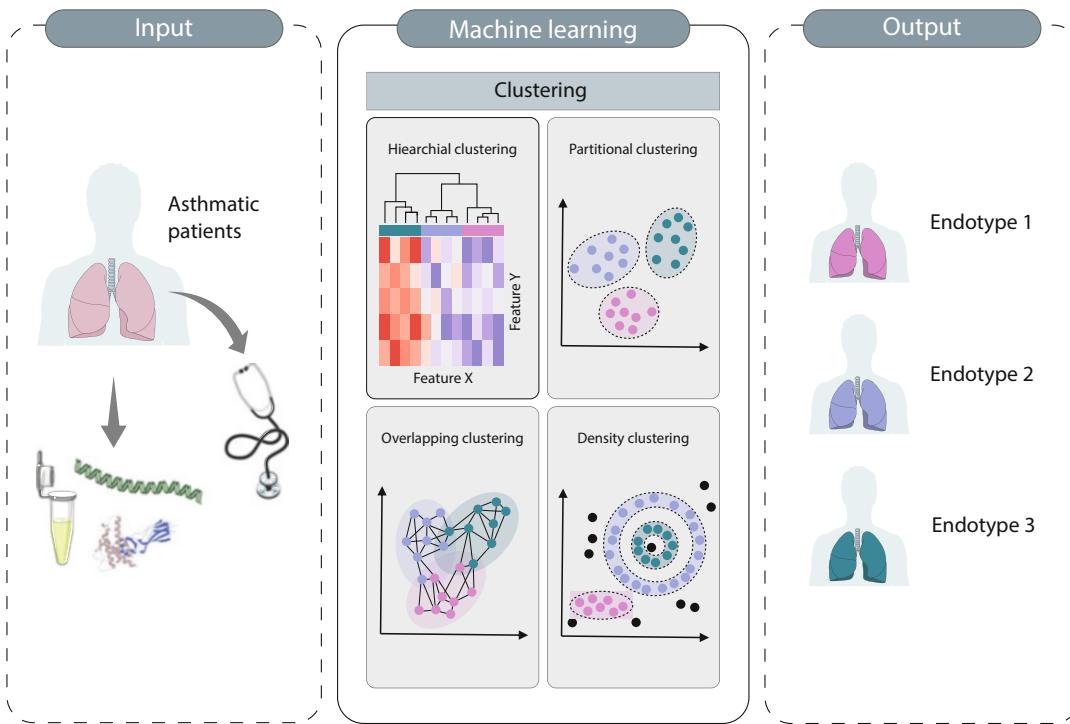


Fig. 3 Example of how ML methods can be used for disease stratification through determination of endotypes. Using clinical data or omics, or both (leftmost column), unsupervised ML-based clustering methods can be applied (middle column) to find patterns for disease stratification. As a result, patients may be categorized into disease

subgroups, such as endotypes, based on molecular features associated with a certain cluster (rightmost column). (This figure was created using vectors from Servier Medical Art, under a Creative Commons Attribution 3.0 Unported License; <http://smart.servier.com>)

on the training data but poorly with new data. In general, small training sets do not suffice to build a model that can generalize well on a much larger sample set. Moreover, datasets often lack information regarding confounders and can incorporate features that are perturbed by technical variation (e.g., batch effects), adding unwanted bias. Thus, a careful study design, including well-characterized patient groups, and methods that ensure high data quality, is necessary for generating reliable results.

Nevertheless, it is possible to solve the problem of small sample sets. One way is to systematically include external data sets in the analysis. For omics studies, external data sets are available in public repositories, such as the ArrayExpress, Gene Expression Omnibus (GEO), GenomeRNAi and dbGAP. Automatizing the search and reuse of publicly available data for ML-driven approaches

is a laborious task, but it is considerably facilitated by the current availability of several omics search engines, which can link omics studies with a similar experimental setup (e.g., same disease, same tissue, similar omics technology, similar clinical phenotype, etc.) [61, 62].

In the near future, more advanced ML techniques will be needed in the field of allergology, and particularly ML models enabling the use of real-time data will play a major role. For example, a recently published pilot study evaluated the feasibility of a deep learning algorithm that provides people with real-time risk assessment of their asthma exacerbation based on the dynamic patterns of indoor air pollutants and their general health condition [63]. Another study investigated the use of deep learning models for the prediction of allergic reactions by using safety event reports across hospitals [64]. For the time being, ML

applications based on real-time data are not available for omics data. Although omics technologies have improved tremendously over the past 30 years, they are still expensive and not readily accessible to most individual laboratories and scientists. Therefore, it is a challenge to build a continuous workflow from omics data to patient monitoring systems. However, within the last few years, new omics-based technologies, including Oxford Nanopore Technologies (ONT, Oxford, UK) [65], have been emerging. These new technologies enable genomics experiments to be performed in minimal facilities and operated in essentially any environment. Moreover, very recently, iGenomics, which is the first iOS application for pre-processing genomics data in a mobile environment, has been released [66]. This emerging new technology will bring us closer to the future of medical tests that utilize real-time genomics data and ML-based methodology for creating fast and reliable diagnostic tools.

In conclusion, the current accumulation of clinical and research-related data will allow for more applications of AI in healthcare and high-performance data-driven medicine. However, many of these applications are still under development, and rarely used in the clinic. Medical professionals will need to understand and adapt to these advances, for better transition of these technologies to the clinic. Finally, any ethical concerns, such as those regarding privacy or representative biases, will need to be addressed.

References

- Murison LB, Brandt EB, Myers JB, Hershey GK. Environmental exposures and mechanisms in allergy and asthma development. *J Clin Invest* [Internet]. 2019;129(4):1504–15. <https://doi.org/10.1172/JCI124612>.
- Strachan DP. Hay fever, hygiene, and household size. *BMJ* [Internet]. 1989 Nov 18 [cited 2021 Feb 24];299(6710):1259–60. <https://www.bmjjournals.com/content/299/6710/1259>
- Rook GAW, Brunet LR. Microbes, immunoregulation, and the gut. *Gut* [Internet]. 2005 Mar 1 [cited 2021 Feb 24];54(3):317–20. <https://gut.bmjjournals.com/content/54/3/317>
- von Hertzen L, Hanski I, Haahtela T. Natural immunity. *EMBO Rep* [Internet]. 2011;12(11):1089–93. <https://doi.org/10.1038/embor.2011.195>.
- Martin SF, Rustemeyer T, Thyssen JP. Recent advances in understanding and managing contact dermatitis. *F1000Res* [Internet]. 2018;7. <https://doi.org/10.12688/f1000research.13499.1>
- Fyhrquist N, Lehto E, Lauerma A. New findings in allergic contact dermatitis. *Curr Opin Allergy Clin Immunol* [Internet]. 2014;14(5):430–5. <https://doi.org/10.1097/ACI.0000000000000092>.
- Borok J, Matiz C, Goldenberg A, Jacob SE. Contact dermatitis in atopic dermatitis children-past, present, and future. *Clin Rev Allergy Immunol* [Internet]. 2019;56(1):86–98. <https://doi.org/10.1007/s12016-018-8711-2>.
- Weidinger S, Beck LA, Bieber T, Kabashima K, Irvine AD. Atopic dermatitis. *Nat Rev Dis Primers* [Internet]. 2018;4(1):1. <https://doi.org/10.1038/s41572-018-0001-z>.
- Lopes JP, Sicherer S. Food allergy: epidemiology, pathogenesis, diagnosis, prevention, and treatment. *Curr Opin Immunol* [Internet]. 2020;66:57–64. <https://doi.org/10.1016/j.coi.2020.03.014>.
- Papi A, Brightling C, Pedersen SE, Reddel HK. Asthma. *Lancet* [Internet]. 2018;391(10122):783–800. [https://doi.org/10.1016/S0140-6736\(17\)33311-1](https://doi.org/10.1016/S0140-6736(17)33311-1).
- Allergic Rhinitis and Its Impact on Asthma: ARIA Workshop Report in Collaboration with the World Health Organization [Internet]. 2001. 188 p. https://books.google.com/books/about/Allergic_Rhinitis_and_Its_Impact_on_Asth.html?hl=&id=51wMzQEACAAJ
- Yang L, Fu J, Zhou Y. Research progress in atopic march. *Front Immunol* [Internet]. 2020;11:1907. <https://doi.org/10.3389/fimmu.2020.01907>
- Mersha TB, Afanador Y, Johansson E, Proper SP, Bernstein JA, Rothenberg ME, et al. Resolving clinical phenotypes into endotypes in allergy: molecular and omics approaches. *Clin Rev Allergy Immunol* [Internet]. 2020. <https://doi.org/10.1007/s12016-020-08787-5>
- Fortino V, Wisgrill L, Werner P, Suomela S, Linder N, Jalonen E, et al. Machine-learning-driven biomarker discovery for the discrimination between allergic and irritant contact dermatitis. *Proc Natl Acad Sci USA* [Internet]. 2020;117(52):33474–85. <https://doi.org/10.1073/pnas.2009192117>.
- Parimbelli E, Marini S, Sacchi L, Bellazzi R. Patient similarity for precision medicine: a systematic review. *J Biomed Inform* [Internet]. 2018;83:87–96. <https://doi.org/10.1016/j.jbi.2018.06.001>.
- Deliu M, Sperrin M, Belgrave D, Custovic A. Identification of asthma subtypes using clustering methodologies. *Palm Ther* [Internet]. 2016;2:19–41. <https://doi.org/10.1007/s41030-016-0017-z>.
- Beaulieu-Jones BK, Greene CS, Pooled Resource Open-Access ALS Clinical Trials Consortium. Semi-

- supervised learning of the electronic health record for phenotype stratification. *J Biomed Inform* [Internet]. 2016;64:168–78. <https://doi.org/10.1016/j.jbi.2016.10.007>.
18. Loftus TJ, Filiberto AC, Li Y, Balch J, Cook AC, Tighe PJ, et al. Decision analysis and reinforcement learning in surgical decision-making. *Surgery* [Internet]. 2020;168(2):253–66. <https://doi.org/10.1016/j.surg.2020.04.049>.
19. Q. Do STAAD. Reinforcement learning framework to identify cause of diseases – predicting asthma attack case. In: 2019 IEEE international conference on Big Data (Big Data) [Internet]. 2019. p. 4829–38. <https://doi.org/10.1109/BigData47090.2019.9006407>
20. Pinart M, Benet M, Annesi-Maesano I, von Berg A, Berdel D, Carlsen KCL, et al. Comorbidity of eczema, rhinitis, and asthma in IgE-sensitised and non-IgE-sensitised children in MeDALL: a population-based cohort study. *Lancet Respir Med* [Internet]. 2014;2(2):131–40. <https://www.sciencedirect.com/science/article/pii/S2213260013702777>
21. Hill DA, Spergel JM. The atopic march: critical evidence and clinical relevance. *Ann Allergy Asthma Immunol* [Internet]. 2018;120(2):131–7. <https://doi.org/10.1016/j.anai.2017.10.037>.
22. Hill DA, Grundmeier RW, Ramos M, Spergel JM. Eosinophilic esophagitis is a late manifestation of the allergic march. *J Allergy Clin Immunol Pract* [Internet]. 2018;6(5):1528–33. <https://doi.org/10.1016/j.jaip.2018.05.010>.
23. Higdon R, Earl RK, Stanberry L, Hudac CM, Montague E, Stewart E, et al. The promise of multi-omics and clinical data integration to identify and target personalized healthcare approaches in autism spectrum disorders. *OMICS* [Internet]. 2015;19(4):197–208. <https://doi.org/10.1089/omi.2015.0020>.
24. Ghosh D, Bernstein JA, Khurana Hershey GK, Rothenberg ME, Mersha TB. Leveraging multilayered “Omics” data for atopic dermatitis: a road map to precision medicine. *Front Immunol* [Internet]. 2018;9:2727. <https://www.frontiersin.org/article/10.3389/fimmu.2018.02727>
25. Stefka AT, Feehley T, Tripathi P, Qiu J, McCoy K, Mazmanian SK, et al. Commensal bacteria protect against food allergen sensitization. *Proc Natl Acad Sci USA* [Internet]. 2014 Aug 22 [cited 2021 Feb 15]. <https://www.pnas.org/content/early/2014/08/21/1412008111>
26. Naval Rivas M, Burton OT, Wise P, Zhang Y-Q, Hobson SA, Garcia Lloret M, et al. A microbiota signature associated with experimental food allergy promotes allergic sensitization and anaphylaxis. *J Allergy Clin Immunol* [Internet]. 2013;131(1):201–12. <https://doi.org/10.1016/j.jaci.2012.10.026>.
27. Park HJ, Lee SW, Hong S. Regulation of allergic immune responses by microbial metabolites. *Immune Netw* [Internet]. 2018;18(1):e15. <https://doi.org/10.4110/in.2018.18.e15>.
28. Ho DSW, Schierding W, Wake M, Saffery R, O’Sullivan J. Machine learning SNP based prediction for precision medicine. *Front Genet* [Internet]. 2019;10:267. <https://www.frontiersin.org/article/10.3389/fgene.2019.00267>
29. Abraham G, Inouye M. Genomic risk prediction of complex human disease and its clinical application. *Curr Opin Genet Dev* [Internet]. 2015;33:10–6. <https://doi.org/10.1016/j.gde.2015.06.005>.
30. Park J-H, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat Genet* [Internet]. 2010;42(7):570–5. <https://doi.org/10.1038/ng.610>.
31. Han Y, Jia Q, Jahani PS, Hurrell BP, Pan C, Huang P, et al. Genome-wide analysis highlights contribution of immune system pathways to the genetic architecture of asthma. *Nat Commun* [Internet]. 2020;11(1):1776. <https://doi.org/10.1038/s41467-020-15649-3>.
32. Hu P, Jiao R, Jin L, Xiong M. Application of causal inference to genomic analysis: advances in methodology. *Front Genet* [Internet]. 2018;9:238. <https://www.frontiersin.org/article/10.3389/fgene.2018.00238/full>
33. Xu M, Tantisira KG, Wu A, Litonjua AA, Chu J-H, Himes BE, et al. Genome Wide Association Study to predict severe asthma exacerbations in children using random forests classifiers. *BMC Med Genet* [Internet]. 2011;12(1):90. <https://doi.org/10.1186/1471-2350-12-90>.
34. Gaudillo J, Rodriguez JJR, Nazareno A, Baltazar LR, Vilela J, Bulalacao R, et al. Machine learning approach to single nucleotide polymorphism-based asthma prediction. *PLoS One* [Internet]. 2019;14(12):e0225574. <https://doi.org/10.1371/journal.pone.0225574>.
35. Ansotegui JJ, Melioli G, Canonica GW, Caraballo L, Villa E, Ebisawa M, et al. IgE allergy diagnostics and other relevant tests in allergy, a World Allergy Organization position paper. *World Allergy Organ J* [Internet]. 2020;13(2):100080. <https://doi.org/10.1016/j.waojou.2019.100080>.
36. Hirota T, Nakayama T, Sato S, Yanagida N, Matsui T, Sugiura S, et al. Association study of childhood food allergy with genome-wide association studies-discovered loci of atopic dermatitis and eosinophilic esophagitis. *J Allergy Clin Immunol* [Internet]. 2017;140(6):1713–6. <https://doi.org/10.1016/j.jaci.2017.05.034>.
37. Prescott S, Allen KJ. Food allergy: riding the second wave of the allergy epidemic. *Pediatr Allergy Immunol* [Internet]. 2011;22(2):155–60. <https://doi.org/10.1111/j.1399-3038.2011.01145.x>.
38. Santos AF, Lack G. Food allergy and anaphylaxis in pediatrics: update 2010–2012. *Pediatr Allergy Immunol* [Internet]. 2012;23(8):698–706. <https://doi.org/10.1111/pai.12025>.
39. Alag A. Machine learning approach yields epigenetic biomarkers of food allergy: a novel 13-gene signature to diagnose clinical reactivity. *PLoS One* [Internet]. 2019 Jun 19 [cited 2021 Feb 15];14(6):e0218253.

- <https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0218253&type=printable>
40. Martino D, Dang T, Sexton-Oates A, Prescott S, Tang MLK, Dharmage S, et al. Blood DNA methylation biomarkers predict clinical reactivity in food-sensitized infants. *J Allergy Clin Immunol* [Internet]. 2015;135(5):1319–28.e12. <https://doi.org/10.1016/j.jaci.2014.12.1933>.
 41. Martino D, Neeland M, Dang T, Cobb J, Ellis J, Barnett A, et al. Epigenetic dysregulation of naïve CD4+ T-cell activation genes in childhood food allergy. *Nat Commun* [Internet]. 2018;9(1):3308. <https://doi.org/10.1038/s41467-018-05608-4>.
 42. Diamandis EP. Cancer biomarkers: can we turn recent failures into success? *J Natl Cancer Inst* [Internet]. 2010 Oct 6 [cited 2021 Feb 16];102(19):1462–7. <https://academic.oup.com/jnci/article/102/19/1462/2515934>
 43. García-Gutiérrez MS, Navarrete F, Sala F, Gasparyan A, Austrich-Olivares A, Manzanares J. Biomarkers in psychiatry: concept, definition, types and relevance to the clinical reality. *Front Psychiatry* [Internet]. 2020;11:432. <https://www.frontiersin.org/article/10.3389/fpsyg.2020.00432>
 44. Torres R, Judson-Torres RL. Research techniques made simple: feature selection for biomarker discovery. *J Invest Dermatol* [Internet]. 2019;139(10):2068–74.e1. <https://doi.org/10.1016/j.jid.2019.07.682>.
 45. Fortino V, Scala G, Greco D. Feature set optimization in biomarker discovery from genome-scale data. *Bioinformatics* [Internet]. 2020;36(11):3393–400. <https://doi.org/10.1093/bioinformatics/btaa144>.
 46. Bieber T, D'Erme AM, Akdis CA, Traidl-Hoffmann C, Lauener R, Schäppi G, et al. Clinical phenotypes and endophenotypes of atopic dermatitis: where are we, and where should we go? *J Allergy Clin Immunol* [Internet]. 2017;139(4S):S58–64. <https://doi.org/10.1016/j.jaci.2017.01.008>.
 47. Werfel T, Allam J-P, Biedermann T, Eyerich K, Gilles S, Guttman-Yassky E, et al. Cellular and molecular immunologic mechanisms in patients with atopic dermatitis. *J Allergy Clin Immunol* [Internet]. 2016;138(2):336–49. <https://doi.org/10.1016/j.jaci.2016.06.010>.
 48. Fyhrquist N, Muirhead G, Prast-Nielsen S, Jeanmougin M, Olah P, Skoog T, et al. Microbe-host interplay in atopic dermatitis and psoriasis. *Nat Commun* [Internet]. 2019;10(1):4703. <https://doi.org/10.1038/s41467-019-12253-y>.
 49. Kuruvilla ME, Lee FE-H, Lee GB. Understanding asthma phenotypes, endotypes, and mechanisms of disease. *Clin Rev Allergy Immunol* [Internet]. 2019;56(2):219–33. <https://doi.org/10.1007/s12016-018-8712-1>.
 50. Fitzpatrick AM, Teague WG, Meyers DA, Peters SP, Li X, Li H, et al. Heterogeneity of severe asthma in childhood: confirmation by cluster analysis of children in the National Institutes of Health/National Heart, Lung, and Blood Institute Severe Asthma Research Program. *J Allergy Clin Immunol* [Internet]. 2011;127(2):382–9.e1–13. <https://doi.org/10.1016/j.jaci.2010.11.015>.
 51. Siroux V, Basagaña X, Boudier A, Pin I, Garcia-Aymerich J, Vesin A, et al. Identifying adult asthma phenotypes using a clustering approach. *Eur Respir J* [Internet]. 2011;38(2):310–7. <https://doi.org/10.1183/09031936.00120810>.
 52. Moore WC, Meyers DA, Wenzel SE, Teague WG, Li H, Li X, et al. Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. *Am J Respir Crit Care Med* [Internet]. 2010;181(4):315–23. <https://doi.org/10.1164/rccm.200906-0896OC>.
 53. Prosperi MCF, Sahiner UM, Belgrave D, Sackesen C, Buchan IE, Simpson A, et al. Challenges in identifying asthma subgroups using unsupervised statistical learning techniques. *Am J Respir Crit Care Med* [Internet]. 2013;188(11):1303–12. <https://doi.org/10.1164/rccm.201304-0694OC>.
 54. Donovan BM, Bastarache L, Turi KN, Zutter MM, Hartert TV. The current state of omics technologies in the clinical management of asthma and allergic diseases. *Ann Allergy Asthma Immunol* [Internet]. 2019;123(6):550–7. <https://www.sciencedirect.com/science/article/pii/S108112061931049X>
 55. Tyler SR, Bunyavanich S. Leveraging -omics for asthma endotyping. *J Allergy Clin Immunol* [Internet]. 2019;144(1):13–23. <https://doi.org/10.1016/j.jaci.2019.05.015>.
 56. Yeh Y-L, Su M-W, Chiang B-L, Yang Y-H, Tsai C-H, Lee YL. Genetic profiles of transcriptomic clusters of childhood asthma determine specific severe subtype. *Clin Exp Allergy* [Internet]. 2018. <https://doi.org/10.1111/cea.13175>.
 57. Sinha A, Desiraju K, Aggarwal K, Kutum R, Roy S, Lodha R, et al. Exhaled breath condensate metabolome clusters for endotype discovery in asthma. *J Transl Med* [Internet]. 2017;15(1):262. <https://doi.org/10.1186/s12967-017-1365-7>.
 58. Nicodemus-Johnson J, Myers RA, Sakabe NJ, Sobreira DR, Hogarth DK, Naureckas ET, et al. DNA methylation in lung cells is associated with asthma endotypes and genetic risk. *JCI Insight* [Internet]. 2016;1(20). <https://doi.org/10.1172/jci.insight.90151>.
 59. Lee-Sarwar KA, Kelly RS, Lasky-Su J, Zeiger RS, O'Connor GT, Sandel MT, et al. Integrative analysis of the intestinal metabolome of childhood asthma. *J Allergy Clin Immunol* [Internet]. 2019;144(2):442–54. <https://doi.org/10.1016/j.jaci.2019.02.032>.
 60. McGeachie MJ, Dahlin A, Qiu W, Croteau-Chonka DC, Savage J, Wu AC, et al. The metabolomics of asthma control: a promising link between genetics and disease: integrative metabolomics of asthma control. *Immun Inflamm Dis* [Internet]. 2015;3(3):224–38. <https://doi.org/10.1002/iid3.61>.
 61. Perez-Riverol Y, Bai M, da Veiga LF, Squizzato S, Park YM, Haug K, et al. Discovering and linking public omics data sets using the Omics Discovery Index. *Nat*

- Biotechnol [Internet]. 2017;35(5):406–9. <https://doi.org/10.1038/nbt.3790>.
62. Chen X, Gururaj AE, Ozyurt B, Liu R, Soysal E, Cohen T, et al. DataMed – an open source discovery index for finding biomedical datasets. *J Am Med Inform Assoc* [Internet]. 2018;25(3):300–8. <https://doi.org/10.1093/jamia/ocx121>.
63. Kim D, Cho S, Tamil L, Song DJ, Seo S. Predicting asthma attacks: effects of indoor PM concentrations on peak expiratory flow rates of asthmatic children. *IEEE Access* [Internet]. 2020;8:8791–7. <https://doi.org/10.1109/ACCESS.2019.2960551>.
64. Yang J, Wang L, Phadke NA, Wickner PG, Mancini CM, Blumenthal KG, et al. Development and validation of a deep learning model for detection of allergic reactions using safety event reports across hospitals. *JAMA Netw Open* [Internet]. 2020;3(11):e2022836. <https://doi.org/10.1001/jamanetworkopen.2020.22836>.
65. Kono N, Arakawa K. Nanopore sequencing: Review of potential applications in functional genomics. *Dev Growth Differ* [Internet]. 2019;61(5):316–26. <https://doi.org/10.1111/dgd.12608>.
66. Palatnick A, Zhou B, Ghedin E, Schatz MC. iGenomics: comprehensive DNA sequence analysis on your Smartphone. *Gigascience* [Internet]. 2020;9(12). <https://doi.org/10.1093/gigascience/giaa138>



Joseph Davids and Hutan Ashrafiyan

Contents

Introduction	1426
Discussion	1426
History	1426
Artificial Intelligence in Diagnostic Haemopathology	1428
Artificial Intelligence in Haem-oncology, Cancer Stem Cells, and Cancer Immunotherapy	1430
AI for Haemo-virology, Haemo-parasitology, and Immune System Studies	1434
AI for Haemorrhage Prediction and Transfusion	1435
AI for Bone Marrow Haematopoietic Stem Cell Transplantation	1435
Point-of-Care Diagnostics, Precision Diagnostics, and Sports Haematology	1436
Future Considerations for AI in Haematology	1437
References	1437

Abstract

Haematology is a field that offers the gateway to life itself. This chapter provides a very brief treatment and discussion of haematology and artificial intelligence, looking at areas where machine learning has captivated the imaginations of haematologists to aid the diagnosis and management of blood-related disorders.

We commence the chapter with a historical overview of AI in haematology and then investigate the subspecialties of haematology where AI has been used to discover new answers to various questions that haematologists have been asking for many decades. We explore decision support systems for diagnostic haemopathology, genetic profiling in haematology, and areas such as immune

J. Davids (✉)

Imperial College London NHS Trust, London, UK

Hamlyn Centre for Robotics and Artificial Intelligence,
Department of Surgery and Cancer, Imperial College
London, London, UK

National Hospital for Neurology and Neurosurgery Queen
Square, London, UK
e-mail: jdavids@ic.ac.uk

H. Ashrafiyan
Department of Surgery and Cancer, Imperial College
London NHS Trust, London, UK

Institute of Global Health Innovation, Imperial College
London, London, UK

Hamlyn Centre for Robotics and Artificial Intelligence,
Department of Surgery and Cancer, Imperial College
London, London, UK
e-mail: h.ashrafiyan@imperial.ac.uk

checkpoint blockade analysis in haem-oncology including AI in leukaemia, lymphoma, and lymphoproliferative disorders, as well as AI for haemo-parasitology and AI for haemo-virology.

We then suggest future areas where AI may impact the treatment of current diseases such as sickle cell anaemia.

Introduction

Blood has signified life throughout the ages, but modern haematology has had to continuously evolve to meet the changing and high-throughput demand for this specialty. For example, our improved and comprehensive understanding of the blood group system has allowed considerable advances in the approach to managing blood transfusion-related disorders. Under the guidance of a haematologist, the modern clinician is now usually involved in transfusion medicine, thus ensuring that the right blood is available for patients when required. Haematologists also provide crucial support for haem-oncology management, stem cell transplantation, and the detection of blood cancers such as leukaemia, etc.

The role of haematologists in the perioperative period is also critical. They ensure that the right crossmatched blood is available so that patients can be adequately transfused intraoperatively, even in an emergency situation, for example, anaemic patients with anticipated high blood loss undergoing complex surgical procedures, patients with pre-existing complex syndromic haematological conditions, or actively haemorrhaging patients. Antigens in the patient's blood that could lead to dangerous adverse transfusion reactions also need to be identified early to circumvent transfusion-related complications.

Similarly, artificial intelligence and its subfacets of deep and machine learning have continued to achieve significant diffusion into most clinical subspecialties including neurology, dermatology, cardiology, rheumatology, genetics, and cardiothoracic surgery etc. All of which have been discussed in detail in various chapters within this book. Haematology is another specialty that is benefitting from the advances in

artificial intelligence and seems to have significantly embraced it for laboratory diagnostics. Haematopathological analysis of benign and malignant disorders of red cell, white cell, platelet, and bone marrow pathologies requires unique methods for accurate diagnostics and is an area that machine learning promises to augment.

This chapter reviews the literature discussing areas where artificial intelligence has been applied to haemato-rheological diagnoses. We review and discuss the subspecialty areas of haematology in which AI has made successful inroads. We then extend the discussion to future areas where we believe haematology could benefit from the multiple facets of artificial intelligence.

Discussion

The exponential growth and explosion of medical data and insight discovery have enabled significant advances to be made in multiple fields including haematology. The era of big data has also led to advances in medical informatics, opening doors for insight discovery from sparse medical data. We begin this chapter discussion with a historical overview of AI as it pertains to haematology and then extend the discussion to areas where we believe AI has impacted haematological diagnosis and treatment, including leukaemia diagnosis, etc.

History

Rosenblatt is quoted as one of the pioneers who presented the modern learning machine paradigm in the 1960s, which detected optical patterns, and that led to the development of the first adaptive artificial neural network called the multiple adaptive linear elements (MADALINE) used for the real-world problem of eliminating phone-line echoes [1, 2]. Other noteworthy projects include Stanford's DENDRAL, from which MYCIN evolved and leveraged heuristic programming for detailed task-specific scientific hypothesis discovery [3]. For details about the history of machine learning and AI, the reader is referred to the early chapters that provide a detailed account of the historical origins of the field. We

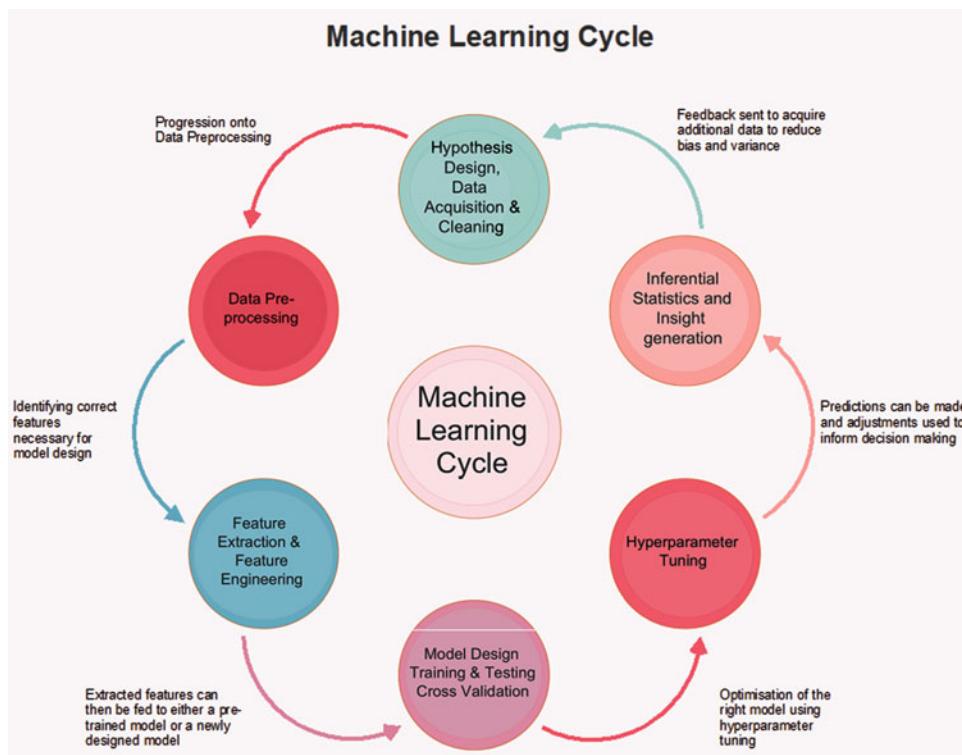


Fig. 1 Illustrates the machine learning cycle. Presented here is one of the many approaches to machine learning, moving counterclockwise from data acquisition to data pre-processing; to feature engineering; to model design, training, and testing; to optimization; and to inference

generation. This is a cyclical process, where additional data from the last step (inference) is utilized to improve the model and thus continue the cyclical nature to the process

do however summarize an example of the modern methodology of how a machine learning pipeline unfolds in Fig. 1. Since 1998, when just about 400 publications were available for artificial neural networks, the field has now seen a tremendous explosion of research interest and adoption into clinical practice with consequent advancements, ranging from deep neural networks to generative adversarial networks. Recent machine learning paradigms such as meta-learning and quantum machine learning have also been developed and have been discussed in various chapters in this book. Nevertheless, this section will look at the history of artificial intelligence in relation to haematology and evolve discussions from there.

Initial applications of machine learning for haematology were linked to clinical haematological laboratory investigations, diagnostics, and decision support as three knowledge-based systems installed in five European hospitals dating back to

1995 [2, 4–9]. These tri-interactive systems were named Professor Petrushka, Professor Fidelio (a heuristic-based diagnostic pattern system, tested on leukaemias, lymphoproliferative disorders, and lymphomas), and Professor Belmonte [2, 7–11]. They were created for peripheral blood interpretation, flow cytometry immunophenotyping, and bone marrow reporting and generated 94% diagnostic accuracy compared with 99% human counterpart on haem-oncology diagnostics [10]. However, this paved the way for intelligent knowledge-based systems in haematology. Apart from accuracy, these systems were also efficient and reduced the turnaround time.

Various studies are summarized in Table 1 below and have looked at machine learning applications for acute myeloid leukaemia, acute promyelocytic leukaemia, aplastic anaemia, and bone marrow disorders. We begin our treatise on diagnostic haemopathology using artificial intelligence.

Table 1 Studies discussing artificial intelligence and machine learning models in acute myeloid leukaemia (AML). *HCST* haematopoietic stem cell transplantation, *TCGA* The Cancer Genome Atlas, *DDx* differential diagnosis, *NGS* next-generation sequencing, *GBT* gradient boosted tree, *ELN* European Leukaemia Net, *HMA*

Author, year	Study size	Application	ML model	Results
Warnat-Herresthal et al. 2019 [30]	12,029 AML, leukaemia, control patients; 102 studies	Transcriptomics-based AML diagnosis and subtype classification	LASSO regression	98% accuracy for AML DDx; ≥ 97% accuracy for AML subclassification (except FAB M7 leukaemia)
Shreve 2019 [31]	3,421 AML patients	Prediction of OS based on 44-gene NGS panel	GBT	C-index of 0.80 compared to ELN C-index of 0.59; individual-level prediction explanations
Lee et al. 2018 [32]	30 patients, 160 drugs	Multi-omic approach to predicting drug sensitivities	Novel model (MERGE)	Identification of gene hubs/networks from TCGA; reproduction of known genomic markers; identification of putative novel drug susceptibilities
Nazha et al. 2019 [33]	433 MDS patients	NGS panel-based prediction of HMA resistance	Recommender algorithm	Stratification of patients with ≥3 mutations into high risk (OS = 14.6 months) and low risk (OS = 22.8 months)
Krug et al. 2010 [34]	2,208 AML patients	CR and mortality prediction from clinical data	Multiple regression	AUROC 0.72 with cytogenetic risk factors; AUROC 0.65 with clinical data alone
Gerstung et al. 2017 [35]	1,540	Risk stratification for HSCT in AML	Novel regression method	C-index of 0.72 for overall survival, individualized risk prediction for HSCT in CR1 versus relapse
Kimura et al. 2019 [36]	3,261 peripheral smears	Differentiation between MDS/aplastic anaemia	CNN	96.2% sensitivity, 100% specificity, 0.99 AUROC for distinguishing MDS/aplastic anaemia
Chandradevan et al. 2019 [37]	10,000 annotated cells	Classification of cells on bone marrow aspirate	CNN	0.98 AUROC classification on non-neoplastic tissue
Shouval et al. 2016 [38]	26,266 AML + HCST patients	Prediction of mortality s/p HCST	Multiple models including naïve Bayes, decision tree, and logistic regression	AUROC up to 0.72; identified key

Artificial Intelligence in Diagnostic Haemopathology

The need to effectively identify abnormalities in peripheral blood for effective clinical decision-making is paramount in clinical practice and serves as one of the backbones of good clinical care. Clinicians need diagnostic haematology to

hypomethylating agent, *AUROC* area under the receiver operating characteristic curve, *CNN* convolutional neural network, *LASSO* least absolute shrinkage and selection operator, *MDS* myelodysplastic syndrome, *CR* complete remission, *OS* overall survival. (Adapted from Radakovich [29])

support, refute, or primarily augment their clinical hypothesis and usually require this information at the bedside following history-taking and examination. Leukocyte count, peripheral blood films, and biochemistry on blood samples can identify infections, haemoglobinopathies, acute and chronic conditions including oncology, and familial and genetic disorders. These tests are crucial

adjuncts to the effective diagnosis of multi-systemic conditions and form the bedrock of modern clinical care.

Use of Neural Networks for Peripheral Blood Film Analysis

A neural network with a 100% diagnostic accuracy was trained on 512 peripheral blood samples (146 normal) [2]. This was then used to perform peripheral blood film analysis on the ADVIA 120, a Mie theory-based blood analyzer that was built as a laser cytometer. The device applies isovolumetric spherling containing sodium dodecyl sulfate and auto-analyzes about 40,000 red blood cells. Automated analysis via neural nets was achieved using unsupervised learning paradigms leveraging concepts of inductive reasoning to create a self-organized inductive map to help classify haematopathological diseases. Preliminary data validating their system on 1,025 samples demonstrated 703 abnormalities ranging from alpha- and beta-thalassemia, with globin variants E, CS, iron deficiency anaemia (IDA) and IDA with thalassemia, and other identifiable haemoglobinopathies. There were 322 normal samples.

Classification of White Blood Cells Using Transfer Learning

Leukocytes are defenders of our immune system and interact with pathogens that enter the body. Using machine learning to classify, identify, and screen these immune defenders for deficits has been presented in the literature. One such study showed that two deep convolutional neural networks showed the best classification performances for white blood cell classification [12]. In this study, the method of two-module transfer learning weighted optimized deformable convolutional neural networks (TWO-DCNN) was proposed for WBC classification, and the performance was validated against classical algorithms commonly employed in machine learning science including support vector machines, standard neural networks, decision trees, and random forest. They also used prebuilt models such as VGG16 and VGG19, inception V3, and the ResNet-50 algorithms. Their best reported performance was a

precision of 95.7% and recall of 95.7%, F1-scores (F1s) of 95.7%, and area under curves (AUCs) of 0.98, for low-resolution images. Although this is impressive, one disadvantage is the fact that the data was a noisy dataset.

Screening Blood Disorders like Thalassemia

Others have leveraged logistic regression, k-nearest neighbors, support vector machine, extreme learning machine and regularized extreme learning machine, AdaBoost, and neural networks for screening purposes in inherited hemoglobinopathies, which are known to arise in 7% of the world's population [2, 13–15]. Erler and others demonstrated thalassemic variant detection in 1995 using a backpropagating artificial neural network using optimized linear and quadratic discriminant functions with reported discriminant efficiencies of 94% [16]. Similarly, others have used haema-chromocytometric analysis on red blood cell, haemoglobin, haematocrit, and mean cell volume parameters that demonstrated a 94% accuracy using a feedforward artificial neural network [2, 17–20].

In another study on 342 patients, classification performance was evaluated with accuracy, sensitivity, f-measure, and specificity parameters using haemoglobin, RBC, HCT, MCV, MCH, MCHC, and RDW with reported accuracy of 96.30% for females, 94.37% for males, and 95.59% for both [13].

The classification of various leukocytes is an area that benefits from automated cytochemistry due to the high throughput of requests for diagnostic investigations. Identifying the right class of leukocytes is dependent upon optical methods of characterization of volume, myeloperoxidase activity, nuclear density, and chromatin content as well as cellular distribution. From this, leukocyte types can be differentiated based on the above as well as their chromatin content and myeloid and lymphoid pattern distribution [2, 21].

A feedforward artificial neural network trained with 84 features was built and used to identify and differentiate normal vs abnormal archetypes in blood and achieved 89% accuracy for Hb E/b-thalassemia [22].

Screening for Polycythaemia Rubra Vera

Decision rule extraction through data mining was used to identify an optimized set of parameters for polycythaemia diagnosis from eight features identified by the gold standard polycythaemia vera study group criteria including gender and haematocrit value. An artificial neural network was able to diagnose polycythemia in 98% of cases [23].

Screening for Pyruvate Kinase Deficiency

Other complex inherited diseases that have benefitted from employing machine learning include the use of binary classification models in metabolomics for pyruvate kinase deficiency on dried blood spots, with an AUC of 0.99 and an accuracy of 94% [24].

Digital Morphological Analysis

Machine learning algorithms have been employed for the morphological analysis of blood smears using convolutional neural networks [8, 25]. Kalman filters for physiological cardiovascular dynamics such as blood pressure and heart rate were also used for diffused optical tomography in oxyhaemoglobin analysis [11].

Artificial Intelligence in Haem-oncology, Cancer Stem Cells, and Cancer Immunotherapy

With the rising global burden of new haematological cancers reaching approximately 18.1 million in 2018, haematological and other cancers affect people of all ages; hence, newer methods of screening, diagnostics, and approaches to treatment have been proposed to aid early identification [26].

Various methods have been adopted to aid this, ranging from haematological bone marrow biopsies, immunological tests, flow cytometric methods, etc. The gold standard haematological test opening the gateway into blood disorder diagnosis for blood cancers is the full blood count (complete blood count). Derangements in full blood counts can provide key diagnostic clues for haematological malignancies.

It is hoped that the application of machine learning and artificial intelligence will reduce the diagnostic conundrum associated with haematological disease, by narrowing down the combinatorial diagnostic challenges and thereby aiding the prediction of haem-oncological disease.

AI for Haem-oncology Screening

In a population-based screening study by Syed-Abdul and colleagues, they applied machine learning algorithms such as artificial neural networks(ANN), stochastic gradient descent, support vector machine, random forest, decision tree, linear models, and logistic regression to the diagnosis of International Classification of Disease-10-coded haematological diseases [26]. These included malignant neoplasms of lymphoid and haematopoietic and related tissue, nutritional anaemia, haemolytic anaemia, aplastic and other anaemia, coagulation defects, purpura and other haemorrhagic conditions, and diseases of blood and blood-forming organs. The ANN with the neural network outperformed the other machine learning models in diagnostic accuracy, precision, recall, and $AUC \pm$ standard deviation as follows: 82.8%, 82.8%, 84.9%, and $93.5\% \pm 2.6$, respectively.

AI for Immune Checkpoint Blockade Discovery

Immune checkpoint blockade refers to the concept of being able to direct the host's immune system against an aggressively dividing tumor cell [27]. The body's own immune system performs the necessary checks to ensure that the immune response to a particular insult does not elicit aggressive immune overactivity. Blocking these checks leads to an uncontrollable immune response in certain individuals and has been found to be effective at fighting off cancers and improving outcomes. These checkpoints are triggered when immune cells, such as T-cell surface proteins, bind to partner cell surface proteins to verify that they are normal. In the case of tumors, the T-cell surface proteins can bind to a tumor cell surface protein, which will have evolved to switch off the checkpoint status to treat the interaction as normal.

Chemotherapeutic checkpoint inhibition drugs therefore allow the immune system to attack the dividing tumor cell. The immune system's T-cells are thus misled into not attacking the tumor cell. So when this checkpoint system inhibiting the "off signal" to the immune system is not triggered, the immune system can attack the tumor cells. Still, sometimes the checkpoint protein inhibitors do not work, and this dampens the response for the patient placed on this particular chemotherapy agent. Two established biomarkers such as programmed death ligand 1 (PD-L1) and tumor mutational burden (TMB) can enable the selection of patients who will benefit most from ICI [27]. Artificial intelligence has been adopted to help predict the response to the chemotherapeutic agent in certain patients with haematological cancers.

Kim et al. built Neopepsee as one such machine-learning-based neoantigen prediction platforms which incorporates nine immunogenicity features to determine immunogenic neoantigens in melanoma and chronic lymphocytic leukaemia [28].

Leukaemia and Lymphoproliferative Disorders

Leukaemia is a blood tissue-forming cancer and a wide-ranging group of disorders with characteristic differences that require appropriate diagnostic protocols to identify. Some forms of leukaemia affect children, whereas others affect adults.

Machine learning has been applied to the diagnosis, subtype classification, treatment, progression monitoring, and other aspects of the disease, which are explored in the subsequent sections.

AI for Acute Myeloid Leukaemia

The models used in leukaemia range from recommender systems to convolutional neural networks. We summarize them in Table 1 below and explore some others in detail [29].

Warnat-Herresthal et al. applied Lasso regression to predict transcriptomic changes in 12,029 acute myeloid leukaemia patients adopting it for subtype classification at 98% accuracy [30].

Convolutional neural networks were used for bone marrow cellular classification and

differentiation between anaplastic anaemia with over 96.2% accuracy [36, 37]. Mortality prediction was achieved at an area under receiver operator curve of 72% using a combination of in silico approaches ranging from naïve Bayes, alternating decision tree, and logistic regression in 26,266 AML and haematopoietic stem cell transplant patients. A summary of the research is presented in [29].

AI for Acute Promyelocytic Leukaemia

Artificial neural networks have again been adopted for the predictive modelling of acute promyelocytic leukaemia. The model was applied to full blood counts to predict APML from non-APML groups with a 95–97% accuracy with a 2.3% false prediction rate [39].

AI for Lymphoma

A semantic network-based artificially intelligent knowledge-based platform capable of describing pathologically aberrant antigen expression in non-Hodgkin's lymphoma diagnosis showed a 97% accuracy [40]. Immunophenotyping is crucial for non-Hodgkin's B-cell lymphoma diagnosis with multiple clinical-pathological entities. Machine learning was applied to a database of 1,465 B-cell non-Hodgkin's lymphoma samples to build 4 artificial intelligence platforms for diagnosis that were able to identify 9 of the most common clinical-pathological entities with an accuracy of 92.68%, sensitivity of 88.5%, and mean specificity of 98.77% [41].

Mantle cell and small lymphocytic lymphomas are matured B-cell lymphomas that are known to have similar immunophenotypic profiles and are usually diagnosed with flow cytometry, which can be ambiguous and thus suboptimal. An artificial intelligence approach to automatically differentiate the two by mean CD20/CD23 fluorescent intensity ratios has been discussed in the literature by Zare and colleagues with a 100% accuracy for mantle cell lymphoma classification and 97% accuracy for small lymphocytic lymphoma [42].

AI in Acute Lymphoblastic Leukaemia

Acute lymphoblastic leukaemia (ALL) is a common childhood cancer with a variety of powerful

risk factor prognosticators of which the most independent is its response to induction chemotherapy according to Bradford et al. and others [43, 44]. Machine learning has also been applied in its capacity as a monitoring tool for patients receiving treatment for ALL. A deep neural network and various other learning algorithms like support vector machines, random forests, and naïve Bayes were trained to differentiate an ALL cell from normal B-lymphocytes. Deep neural networks were then used to extract discriminating bright- and dark-field imaging features ignoring fluorescence approaches on antibody-labelled diagnostic and on-treatment bone marrow samples that had been labelled. Analysis showed accuracy of 88%. Other approaches included feature extraction and dimensionality reduction leveraging principal component analysis and t-distributed stochastic neighbor embedding (t-SNE) for visualization [44].

AI in Multiple Myeloma and Cancer Stem Cells

Epidemiologically, multiple myeloma has a median age of diagnosis of 70 years. The diagnosis is heralded by identifying the clonal expansion of plasma cells and includes monoclonal gammopathy of uncertain significance [45].

Single-cell cancerous subclones that are usually suppressed by current treatments can rapidly switch from dormant to dominant subclones in the pathophysiological course of multiple myeloma, and this subclone switching can lead to cancer drug resistance. Recently, advancements in technology have evolved to track the subclone formation [45].

AI in Neoplastic Bone Marrow Disease

An artificially intelligent knowledge-based workstation built for peripheral blood analysis, flow cytometry immunophenotyping, and bone marrow morphology, linked together by a relational database, was trained to classify neoplastic bone marrow pathology in 526 cases. There was agreement with the expert in 97.9% of cases ($n = 515$ cases) with 11 diagnostic errors with minimal misclassifications [46]. There was agreement

with the gold standard for diagnosis in 87.8% of cases.

An unsupervised clustering algorithm was used to differentiate between normal and malignant human bone marrow on classical flow cytometry data using the flow self-organizing map, an R-solution applied together with a Kaluza® software platform [47, 48].

AI for Lymphoproliferative Disorders

In the detection of basket cells in lymphoproliferative disorders, artificial intelligence has been used for the identification of degenerate lymphocytes in this condition. These degenerate cells, called smudge cells, were identified using Cellavision. An artificial intelligence platform called the AI-Heme-1 was built by Barouqa and colleagues for the automated detection of basket cells. They created an in vitro model to form NETs in blood and generated a library of their morphological changes at various maturation stages and then correlated their presence to infections in the absence of leukocytosis [49].

Indolent B-cell lymphoproliferative disorders (BLPDs) include chronic lymphocytic leukaemia (CLL), hairy cell leukaemia (HCL), follicular lymphoma (FL), splenic marginal zone lymphoma (SMZL), nodal marginal zone lymphoma (NMZL), mucosa-associated lymphoid tissue (MALT) lymphoma, lymphoplasmacytic lymphoma/Waldenström's macroglobulinemia (LPL/WM), and other unclassified chronic BLPDs. By contrast, the aggressive B-cell lymphoproliferative disorders include diffuse large B-cell lymphoma (DLBCL), Burkitt lymphoma (BL), mantle cell lymphoma (MCL), and B-lymphoblastic lymphoma (B-LBL) [50]. Yi et al. initially applied principal component analysis on 15 T-cell immune signatures generated by flow cytometry, but this was not sufficient to efficiently separate aggressive vs indolent groups due to the heterogeneity of the cellular groups [50]. However, a random forest machine learning model was also used to model the rank of T-cell immune signatures, which showed dysregulation and best performance on binary classification of all indolent B-cell lymphoproliferative disorder patients [50].

AI for Cancer Immunotherapy

In immunotherapy, T-cell distribution and subset T-cell counts demonstrate the effectiveness of the immune system at combating cancer as well as gauging response to therapy. An artificially intelligent immune cell analyzer for T-cell subsets called ImmuCellAI has been developed to help achieve this [51]. It works using a gene-set signature-based method and can identify T-cells, dendritic cells, cytotoxic T-cells, and gamma-delta T-cells.

AI for Haematological Gene Profiling and Molecular Sequencing

Acute leukaemia was used as a template to categorize cancers and stratify them into various classes [2, 52]. Using DNA microarrays, thousands of genes can be identified and screened for pathological variants. Signal-to-noise feature extraction was then used to identify and analyze large gene datasets. Unsupervised and supervised learning methods included self-organizing maps, hierarchical and probabilistic clustering, support vector machines, and k-nearest neighbors [2, 53].

Decision support in systems for oncogenomic analysis has also been developed including the Oncompass calculator – a self-learning AI that is designed to learn from a big dataset and analyze the complex case management of various patients with haem-oncological genomic diseases [6].

For chemically induced DNA damage, the FlowSight® imaging cytometry platform applies machine learning to imaging flow cytometry to access DNA damage via a cytokinesis block micronucleus (CBMN) assay [54]. In their work, Verma and colleagues rapidly imaged and collected a population of about ~20,000 bright-field cells together with DRAQ5™ fluorescent stained nuclei/MN in under 10 min, and in-focus cells were isolated using IDEAS® software. Mono-, bi-, tri-, and tetra-nucleated cells were scored and classified with a mention of supervised machine learning algorithms; however, the specific machine learning algorithms were not clearly identified and are likely proprietary. However, they did mention toxicological modelling using exponential and Hill nested model families.

Other areas that have had machine learning augmentation include single-cell gene expression disease profiling, high-throughput gene profiling, and single-cell RNA transcriptomics for haematopoietic stem cells, where tools such as the hscScores have been scientifically developed with translational considerations [55, 56].

AI for Cancer Disease Progression Monitoring

In their work studying genome-wide expression of chemo-resistant genes noted to cause minimal residual disease for acute myeloid leukaemia cells in 157 patients, Coustan-Smith and colleagues used machine learning-related dimensionality reduction techniques to identify 22 (CD9, CD18, CD25, CD32, CD44, CD47, CD52, CD54, CD59, CD64, CD68, CD86, CD93, CD96, CD97, CD99, CD123, CD200, CD300a/c, CD366, CD371, and CX3CR1) aberrantly expressed markers of minimal residual disease [57]. The machine learning models used included the t-distributed stochastic neighbor embedding (t-SNE) for visualization.

Gaussian mixture models have also been used for minimal residual disease monitoring of acute lymphoblastic leukaemia trained on multi-parameter flow cytometry data from 337 bone marrow samples collected at day 15 of induction therapy [58].

Other machine learning approaches such as support vector machines have been applied on newer technologies of stain-free classification of cancer cells and for blood cell interferometric phase microscopy, which measures the refractive index of the cells [59]. Nassim and colleagues used support vector machines on holographic interferometry to differentiate blood cells from colonic cancer cells. Peripheral blood flow cytometry for the evaluation of circulating cancer cells and monitoring with machine learning models such as logistic regression (100% sensitivity and 54% specificity (AUC = 0.919)) and decision trees (decision tree model achieved 98% sensitivity and 65% specificity (AUC = 0.906)) have also been discussed [60].

AI for Haemo-virology, Haemo-parasitology, and Immune System Studies

AI in COVID-19 Haemo-virology

Our immune system is a sophisticated and well-evolved system that provides the body with support and protection against various pathogens while also distinguishing between foreign substances and the body itself. It is divided into innate immunity, which is the body's first line of defense present from birth, and adaptive immunity. With no memory, the innate immune system encompasses physical and chemical as well as cellular barriers like phagocytic cells and the complement system, whereas the adaptive immune system has memory and is capable of antigen detection, recognition and processing, and the destruction of a previously encountered antigen. However, viruses have evolved to sometimes circumvent this process of detection, as we learnt from the COVID-19 pandemic where there was apparent dysfunction of the adaptive immune system.

COVID-19 spread worldwide causing a global pandemic. It is characterized by acute respiratory distress, multiple organ failure, haematological manifestations, neuro-olfactory disturbances, coagulopathic manifestation, and death. During the COVID-19 pandemic, newer methods of artificial intelligence were used to diagnose and predict the outcome of COVID-19 from blood results. Chung and colleagues used deep neural network, ensemble-based methods, and random forest models called EDRnet [61]. This accurately predicted mortality from blood biomarkers, patient's age, and gender information with a high sensitivity (100%), specificity (91%), and accuracy (92%). A diagnostic web platform was built to facilitate mortality prediction.

AI in HIV and AIDS Haemo-virology

The human immunodeficiency virus (HIV-1) is a global, sexually transmitted disease that has significant repercussions for the patient and society. Machine learning has been suggested to aid in diagnostics and AIDS progression analysis. FloReMI is a method of flow density regression that has been designed to aid prediction of time to

progression of HIV to AIDS using random survival forests [62].

The detection of neutralizing antibodies for effective protection against the virus is also a subject of intense research and an area of current research focus [63]. One study on broadly neutralizing antibodies for the HIV-1 virus aimed to resolve the issue of significant epitope variation that occurs in viral strains, which could be resistant to anti-retroviral drugs. Predicting these epitope variations could enable the identification of resistant strains and thus facilitate drug engineering to combat against drug-resistant strains. Yu et al. designed a novel "Bayesian machine-learning model that uses information from the HIV-1 envelope protein sequences and foremost approximated glycan occupancy information as variables to quantitatively predict the half-maximal inhibitory concentrations (IC₅₀) of 126 neutralizing antibodies against a variety of cross clade viruses" [63]. The model was subsequently applied to peripheral blood and showed that the predicted broadly neutralizing antibody could achieve 100% neutralization and was able to map the degree of neutralizing resistance.

AI in Haemo-parasitology

Research into machine learning for hemo-parasitological diseases such as malaria and dengue fever has also increased in recent decades. Immunochromatography for plasmodium detection on blood film is a diagnostic approach for laboratory malaria and dengue detection [64]. Algorithms have been developed to diagnose malaria and also distinguish between malaria and dengue fever and other febrile conditions [65]. Previously, this was achieved using percentage of lymphocytes and their standard deviation of conductivity on the LH750 with reported area under the curve (AUC) of 0.893 and evaluated AUCs of 0.931 [66]. However, algorithms such as support vector machine platforms have been explored by HORIBA for malaria classification [67, 68].

In another pediatric cross-sectional study, machine learning was used to predict the stimulation levels of two phenotypes of T-cell activations (T-helper cells and regulatory T-cells)

[65]. Predictive analytics for distinguishing the levels of these various phenotypes between symptomatic malaria patients and asymptomatic patients showed a sensitivity of 86% and a specificity of 94%.

Another area reported is the detection of leishmaniasis and Chagas disease and the need for repeated confirmatory serological assays to identify nonnegative results for blood bank screening and to also identify the degree of agreement between methods of diagnosis [69]. Machine learning has also been applied within this area using discriminant analysis to show higher levels of agreement between (1) the multiplexed flow cytometry anti-T. cruzi/Leishmania IgG1 serology/FC-TRIPLEX Chagas/Leish IgG1 and (2) enzyme-linked immunosorbent assay (ELISA)/immunofluorescent consensus criterion (EIA/IIF). The reported Kappa index (Kappa index = 0.811) and Spearman correlation ($r = 0.6171$; $p < 0.0001$) suggested strong agreement between both techniques, but a superior AUC of 0.91 along with sensitivity and specificity (91% and 90%, respectively) in favor of FC-TRIPLEX [69].

AI for Haemorrhage Prediction and Transfusion

AI for Transfusion and Blood Quality Haemo-diagnostics

Transfusion medicine facilitates the haematological treatment of acute hemorrhagic disorders through valid and safe blood transfusion practices that include grouping, screening, and crossmatching to ensure blood is ready for specialty services including surgery.

Significant harm can befall a patient if transfusion quality assessments are skipped or taken for granted, with death as the endpoint. Blood antigen detection is also paramount in identifying other rarer blood group antigens in the sample that could lead to fatal blood transfusion reactions, and these also need to be identified and mitigated. The blood to be transfused must be assessed for its quality, and machine learning has found use cases within this area as well. As blood can undergo

rapid and progressive vitiation due to long-term storage, there is a need to overcome the challenge associated with at times suboptimal and subjective labelling/identification performed by humans as detectors of degraded blood. Algorithmically, deep learning has been proposed to characterize image flow cytometry for life-saving blood [70].

Similar deep learning approaches like the BMSNet, which leverages You Only Look Once (YOLO) and convolutional neural network architectures, have been used for the quality assessment of 122 bone marrow smears containing 17,319 annotated cells to better identify varying disease morphology [71].

Other areas such as scoring systems for upper gastrointestinal hemorrhage prediction have been augmented through gradient boosting machine learning algorithms with reported AUCs of 0.91 [72].

AI for Postpartum Haemorrhage Prediction

Postpartum hemorrhage can be associated with maternal mortality in both developed and developing countries. Machine learning has been used to predict postpartum hemorrhage, identifying 55 candidate risk factors that were studied to predict maternal risk of postpartum hemorrhage [73]. In a patient study of 152,279 assessed births, 7,279 patients with postpartum hemorrhage were identified using logistic regression, Lasso regression, random forests, and gradient boosting. The models were able to also predict postpartum hemorrhage risk to a significantly high degree based on reported C-statistics ranging from 87% for logistic and Lasso regressions to 93% for the extreme gradient boosting model, which was reported as the model with the best discriminative ability [73].

AI for Bone Marrow Haematopoietic Stem Cell Transplantation

Within the expanding field of haematopoietic stem cell transplantation, machine learning is also evolving rapidly. Sixty thousand procedures of haematopoietic stem cell transplants are

performed worldwide, with an estimated half a million survivors [1]. The process carries a heavy emotional burden involving donor-recipient pair selection, where suboptimally screened human leukocyte antigen locus mismatches could prove detrimental to patient safety and decrease 1-year survival from 53% to 43% in optimally matched pairs [74]. Graft-versus-host disease becomes a problem for donor-recipient mismatching, which can lead to transplant failure.

Machine learning models that have been used for pre-transplantation screening to optimize donor-recipient pairing include random forest and logistic regression used to identify amino-acid substitutions that can lead to mismatches [75]. While these algorithms have been promising during pre-validation, clinical validation has presented with failures. Post-transplantation usually presents with uncertainty for both donors and recipients with prognosis/complications such as graft-versus-host disease and others that influence morbidity or mortality [76]. Machine learning has been applied to identify these complications in big data projects such as the Acute Leukaemia Registry established by the European Group for Bone Marrow Transplantation (AL-EBMT) project – a 2015 initiative using alternating decision tree (ADT) analysis on 28,236 patient characteristics, which pursues multiple paths to predict 100-day mortality following allogeneic bone marrow transplantation for acute leukaemia (AUC, 0.702-ADT vs 0.646-EBMT score; $P = 3 \cdot 10^{-18}$) [38].

AI in Autoimmunity

For a comprehensive review in this topic, the reader is directed to the work by Stafford and colleagues, but it will not be a focus in this chapter as these are very organ-specific and will need a separate dedicated chapter [77]. Reported machine learning approaches to autoimmune disorders such as multiple sclerosis, rheumatoid and psoriatic arthritis, systemic sclerosis, autoimmune liver diseases, coeliac disease, inflammatory bowel disease, and type 1 diabetes have used models including support vector machines, variations of random forest, neural networks, decision trees, natural language processing, and other hybrid models [77].

Point-of-Care Diagnostics, Precision Diagnostics, and Sports Haematology

AI in Point-of-Care and Precision Haematological Diagnostics

This section is specific to haematological diagnostic tests for blood sugar monitoring, hemorrhage and coagulopathy, and septic screening, which need to be rapid for immediate decision support. This can be achieved using direct antigen detection from a very small sample of blood, but devices must be calibrated to reflect the true laboratory values. The principles remain the same as for any diagnostic test; it must be accurate and sensitive, safe to administer, cheap and durable to run, and relatively easy to administer, and in an ideal world it must also be painless or delivered with minimal discomfort to the patient.

In order to fulfill the above criteria, the use of rapid point-of-care diagnostic tests became even more necessary with the recent COVID-19 pandemic, which was also characterized by coagulopathy as well as lymphocytic and platelet count derangements. Social distancing measures and clinician safety meant this became even more of a necessity [78, 79].

HemoScreen is one such technology for point-of-care diagnostics leveraging machine learning algorithms for hemo-cellular-based characterization of blood samples on the operating principles of image analysis, microfluidics, and viscoelasticity [78]. HemoScreen correlates well with laser and impedance methods as well [78, 79].

Another area where precision diagnostics matters is in the oncological disease detection of haematological malignancies. Here too, Watson for Genomics is an exemplar IBM platform that has been designed with cognitive computing in mind to offer precision diagnostics using artificial intelligence as reviewed by Yokohama and colleagues [80].

Advances in mHealth and other areas to facilitate point-of-care testing and result processing and delivery have also seen integration with AI-based platforms that have been developed for diabetes and other chronic disease management. However, this will not be a focus here, and the

reader is thus directed to the relevant chapters to supplement their knowledge.

AI in Sports Haematology

Doping scandals have dominated sports headlines over several decades, as intelligently designed performance-enhancing drugs continue to provide athletes with an unfair advantage. Sadly, more sophisticated methods have been adopted to evade haematological tests and may already be taking advantage of machine learning and artificial intelligence methods for drug design and new ways for athletes to avoid detection. There is therefore a global need to tackle this issue with artificial intelligence, big data, and other approaches. One such International Sports Federation-funded project is the Artificial Intelligence Evoked Target Testing (A.R.I.E.T.T.A) project, which is a platform that aims to develop haematological and performance profile analyses to detect abnormal patterns that can isolate an individual for targeted testing [81].

Future Considerations for AI in Haematology

In this short section, we consider areas that will no longer seem challenging in the coming decades in terms of innovation and where AI and haematology will benefit each other.

Diseases like sickle cell anaemia may be an area that will benefit from genetic engineering to ameliorate crises and will be within our reach in the coming decades. Combining machine learning with CRISPR-like technologies could advance the potential for genetically editing out haematological diseases from the genome and will also be useful for genetically augmenting bone marrow and stem cell transplantation therapies to reduce graft-versus-host disease, etc. An example is in HLA haplotyping, which would benefit from better AI algorithms to augment CRISPR-like technologies in screening for loci sequences that can be edited out or into the genome to facilitate and improve better donor-recipient matching (we may even do without the need for matching in the future). The reader is

directed to other relevant chapters for further inspiration to aid their own knowledge discovery.

Nano-robotics in combination with genetic engineering and machine learning will help us to manage rarer genetic blood disorders with curative outcomes. However, significant advances in our understanding and appreciation for the combinatorial differences in rheological characteristics of an individual's blood must be carefully considered and appreciated.

References

- Muhsen IN, ElHassan T, Hashmi SK. Artificial intelligence approaches in hematopoietic cell transplantation: a review of the current status and future directions. *Turk J Haematol.* 2018;35(3):152–7.
- Zini G. Artificial intelligence in hematology. *Hematology.* 2005;10(5):393–400.
- Lindsay R, Buchanan BG, Feigenbaum E, Lederberg J. DENDRAL: a case study of the first expert system for scientific hypothesis formation. *Artif Intell.* 1993;61:209–61.
- Schachter R, Lutter D, Knollmüller P, Tomé AM, Theis FJ, Schmitz G, et al. Knowledge-based gene expression classification via matrix factorization. *Bioinformatics.* 2008;24(15):1688–97.
- Tsien CL. Event discovery in medical time-series data. *Proc AMIA Symp.* 2000;858–62.
- István Vályi-Nagy IP. Development and national roll-out of electronic decision support systems using artificial intelligence in the field of onco-hematology. *Magy Onkol.* 2019;63(4):275–80.
- Diamond LW, Mishka VG, Seal AH, Nguyen DT. A clinical database as a component of a diagnostic hematology workstation. *Proc Annu Symp Comput Appl Med Care.* 1994;298–302.
- Diamond LW, Nguyen DT, Andreeff M, Maiese RL, Braylan RC. A knowledge-based system for the interpretation of flow cytometry data in leukemias and lymphomas. *Cytometry.* 1994;17(3):266–73.
- Diamond LW, Nguyen DT, Lima M, Simón R, Aoûtka SB. A comprehensive knowledge-based system for laboratory hematology. *Comput Methods Prog Biomed.* 1997;54(1–2):69–76.
- Diamond LW, Mishka VG, Seal AH, Nguyen DT. Multiparameter interpretative reporting in diagnostic laboratory hematology. *Int J Biomed Comput.* 1994;37(3):211–24.
- Diamond SG, Huppert TJ, Kolehmainen V, Franceschini MA, Kaipio JP, Arridge SR, et al. Physiological system identification with the Kalman filter in diffuse optical tomography. *Med Image Comput Comput Assist Interv.* 2005;8(Pt 2):649–56.

12. Yao X, Sun K, Bu X, Zhao C, Jin Y. Classification of white blood cells using weighted optimized deformable convolutional neural networks. *Artif Cells Nanomed Biotechnol.* 2021;49(1):147–55.
13. Çil B, Ayyıldız H, Tunçer T. Discrimination of β-thalassemia and iron deficiency anemia through extreme learning machine and regularized extreme learning machine based decision support system. *Med Hypotheses.* 2020;138:109611.
14. Kabootarizadeh L, Jamshidnezhad A, Koohmreh Z. Differential diagnosis of iron-deficiency anemia from β-thalassemia trait using an intelligent model in comparison with discriminant indexes. *Acta Inform Med.* 2019;27(2):78–84.
15. Zhang Y, Han Z, Gao Q, Bai X, Zhang C, Hou H. Prediction of K562 cells functional inhibitors based on machine learning approaches. *Curr Pharm Des.* 2019;25(40):4296–302.
16. Erler BSVP, Lee S. Superiority of neural networks over discriminant functions for thalassemia minor screening of red blood cell microcytosis. *Arch Pathol Lab Med.* 1995;119(4):350–4. PMID: 7726727
17. d’Onofrio G, Zini G, Ricerca BM, Mancini S, Mango G. Automated measurement of red blood cell microcytosis and hypochromia in iron deficiency and b-thalassemia trait. *Arch Pathol Lab Med.* 1992;116:84–9.
18. Birndorf RI, Pentecost JO, Coakley JR, Spackman KA. An expert system to diagnose anemia and report results directly on hematology forms. *Comput Biomed Res.* 1996;29:16–26.
19. d’Onofrio GZG. Diagnostic value of peroxidase and sizeparameters from a new hematological analyzer. *Hema.* 1998;238–39, Proceedings of XXII Congress of ISH. 1998.
20. Amendolia SR, Brunetti A, Carta P, Cossu G, Ganadu ML, Golosio B, Mura GM, Pirastri MG. A real-time classification system of thalassemic pathologies based on artificial neural networks. *Med Decis Mak.* 2002;22:18–26.
21. d’Onofrio GZG. Morphology of the blood. Oxford: Butterworth Heinemann; 1998. In Zini G, Hematology 2005;10(5):393–400. <https://doi.org/10.1080/10245330410001727055>.
22. Zini G, d’Onofrio G. Neural network in hematological malignancies. *Clin Chim Acta.* 2003;333:195–201.
23. Kantardzic M, Djulbegovic B, Hamdan H. A data-mining approach to improving polycythemia vera diagnosis. *Comput Ind Eng Archiv.* 2002;43:765–73.
24. Van Dooijeweert B, Broeks MH, Verhoeven-Duif NM, Van Beers EJ, Nieuwenhuis EES, Van Solinge WW, et al. Untargeted metabolic profiling in dried blood spots identifies disease fingerprint for pyruvate kinase deficiency. *Haematologica.* 2020; Online ahead of print.
25. Ohsaka A. Artificial intelligence (AI) and hematological diseases: establishment of a peripheral blood convolutional neural network (CNN)-based digital morphology analysis system. *Rinsho Ketsueki.* 2020;61(5):564–9.
26. Syed-Abdul S, Firdani RP, Chung HJ, Uddin M, Hur M, Park JH, et al. Artificial intelligence based models for screening of hematologic malignancies using cell population data. *Sci Rep.* 2020;10(1):4583.
27. Huemer F, Leisch M, Geisberger R, Melchardt T, Rinnerthaler G, Zaborsky N, et al. Combination strategies for immune-checkpoint blockade and response prediction by artificial intelligence. *Int J Mol Sci.* 2020;21(8):2856.
28. Kim S, Kim HS, Kim E, Lee MG, Shin EC, Paik S, Kim S. Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information. *Ann Oncol.* 2018;29:1030–6. <https://doi.org/10.1093/annonc/mdy022>.
29. Radakovich N, Cortese M, Nazha A. Acute myeloid leukemia and artificial intelligence, algorithms and new scores. *Best Pract Res Clin Haematol.* 2020;33(3):101192.
30. Warnat-Herresthal S, Perrakis K, Taschner B, Becker M, Baßler K, Beyer M, et al.. Scalable prediction of acute myeloid leukemia using high-dimensional machine learning and blood transcriptomics. *iScience* 2019 December 18 [cited 2020 Mar 4];23(1).
31. Shreve J, et al. A personalized prediction model to risk stratify patients with acute myeloid leukemia (AML) using artificial intelligence. *Blood* 2019;134 (Supplement_1):2091. <https://doi.org/10.1182/blood-2019-128066>.
32. Lee S-I CS, Logsdon BA, Lundberg SM, Martins TJ, Oehler VG, et al. A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. *Nat Commun.* 2018;9(1):42.
33. Nazha ASM, Bejar R, Rauh MJ, Othus M, Komrokji RS, et al. Genomic biomarkers to predict resistance to hypomethylating agents in patients with myelodysplastic syndromes using artificial intelligence. *JCO Precis Oncol.* 2019;3:1–11.
34. Krug U, Röllig C, Koschmieder A, Heinecke A, Sauerland MC, Schaich M, et al. Complete remission and early death after intensive chemotherapy in patients aged 60 years or older with acute myeloid leukaemia: a web-based application for prediction of outcomes. *Lancet Lond Engl.* 2010;376(9757):2000–8.
35. Gerstung M, Papaemmanuil E, Martincorena I, Bullinger L, Gaidzik VI, Paschka P, et al. Precision oncology for acute myeloid leukemia using a knowledge bank approach. *Nat Genet.* 2017;49(3):332–40.
36. Kimura K, Tabe Y, Ai T, Takehara I, Fukuda H, Takahashi H, et al. A novel automated image analysis system using deep convolutional neural networks can assist to differentiate MDS and AA. *Sci Rep.* 2019;9(1):13385.
37. Chandrasevan R, Aljudi AA, Drumheller BR, Kunanthaseelan N, Amgad M, Gutman DA, et al. Machine-based detection and classification for bone marrow aspirate differential counts: initial

- development focusing on nonneoplastic cells. *Lab Investig.* 2020;100:98.
38. Shouval R, Labopin M, Bondi O, Mishan-Shamay H, Shimoni A, Ciceri F, Esteve J, Giebel S, Gorin NC, Schmid C, Polge E, Aljurf M, Kroger N, Craddock C, Bacigalupo A, Cornelissen JJ, Baron F, Unger R, Nagler A, Mohty M. Prediction of allogeneic hematopoietic stem-cell transplantation mortality 100 days after transplantation using a machine learning algorithm: a European Group for blood and marrow transplantation acute leukemia working party retrospective data mining study. *J Clin Oncol.* 33(28):3144–51. <https://doi.org/10.1200/JCO2014591339>. Epub 2015 Aug 3 PMID: 26240227
39. Haider RZ, Ujjan IU, Shamsi TS. Cell population data-driven acute promyelocytic leukemia flagging through artificial neural network predictive modeling. *Transl Oncol.* 2020;13(1):11–6.
40. Thews O, Thews A, Huber C, Vaupel P. Computer-assisted interpretation of flow cytometry data in hematology. *Cytometry.* 1996;23(2):140–9. <https://doi.org/10.1002/cyto.990230202>.
41. Gaidano V, Tenace V, Santoro N, Varvello S, Cignetti A, Prato G, et al. A clinically applicable approach to the classification of b-cell non-hodgkin lymphomas with flow cytometry and machine learning. *Cancers (Basel).* 2020;12(6):1684.
42. Zare H, Bashashati A, Kridel R, Aghaeepour N, Haffari G, Connors JM, et al. Automated analysis of multidimensional flow cytometry data improves diagnostic accuracy between mantle cell lymphoma and small lymphocytic lymphoma. *Am J Clin Pathol.* 2012;137(1):75–85.
43. Bradstock K, Janossy G, Tidman N, et al. Immunological monitoring of residual disease in treated thymic acute lymphoblastic leukaemia. *Leuk Res.* 1981;5 (4–5):301–9.
44. Doan M, Case M, Masic D, Hennig H, McQuin C, Caicedo J, et al. Label-free leukemia monitoring by computer vision. *Cytometry A.* 2020;97(4):407–14.
45. Lee LX, Li SC. Hunting down the dominating subclone of cancer stem cells as a potential new therapeutic target in multiple myeloma: an artificial intelligence perspective. *World J Stem Cells.* 2020;12(8):706–20.
46. Nguyen D, Diamond LW, Cherubino P, Koala WB, Imbert M, Andreeff M. A diagnostic workstation for neoplastic bone marrow diseases: evaluation on 526 cases. *Medinfo.* 1995;8(Pt 1):771–5.
47. Lacombe F, Lechevalier N, Vial JP, Béné MC. An R-derived FlowSOM process to analyze unsupervised clustering of normal and malignant human bone marrow classical flow cytometry data. *Cytometry A.* 2019;95(11):1191–7.
48. Duetz C, Bachas C, Westers TM, van de Loosdrecht AA. Computational analysis of flow cytometry data in hematological malignancies: future clinical practice? *Curr Opin Oncol.* 2020;32(2):162–9.
49. Barouqa M, et al. Neutrophilic extracellular traps (NETs); A subset of smudge cells identifiable by peripheral smear autoanalyzers in the rising era of artificial intelligence. *Am J Clin Pathol.* 2020;154: S10–S11. <https://doi.org/10.1093/ajcp/aqaa137.018>.
50. Yi S, Zhang Y, Xiong W, Chen W, Hou Z, Yang Y, et al. Prominent immune signatures of T cells are specifically associated with indolent B-cell lymphoproliferative disorders and predict prognosis. *Clin Transl Immunol.* 2020;9(1):e01105.
51. Miao YR, Zhang Q, Lei Q, Luo M, Xie GY, Wang H, et al. ImmuCellAI: a unique method for comprehensive T-cell subsets abundance prediction and its application in cancer immunotherapy. *Adv Sci (Weinh).* 2020;7(7): 1902880.
52. Golub TR, Sloim D, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science.* 1999;286:531–7.
53. Beale R, Jackson T. An introduction. London. in Zini G *Hematology* 2005;10(5):393–400. <https://doi.org/10.1080/10245330410001727055>.
54. Verma JR, Harte DSG, Shah UK, Summers H, Thornton CA, Doak SH, et al. Investigating FlowSight® imaging flow cytometry as a platform to assess chemically induced micronuclei using human lymphoblastoid cells in vitro. *Mutagenesis.* 2018;33 (4):283–9.
55. Chlis NK, Rausch L, Brocker T, Kranich J, Theis FJ. Predicting single-cell gene expression profiles of imaging flow cytometry data with machine learning. *Nucleic Acids Res.* 2020;48(20):11335–46.
56. Hamey FK, Göttgens B. Machine learning predicts putative hematopoietic stem cells within large single-cell transcriptomics data sets. *Exp Hematol.* 2019;78: 11–20.
57. Coustan-Smith E, Song G, Shurtleff S, Yeoh AE, Chng WJ, Chen SP, et al. Universal monitoring of minimal residual disease in acute myeloid leukemia. *JCI Insight.* 2018;3(9):e98561.
58. Reiter M, Diem M, Schumich A, Maurer-Granofszky M, Karawajew L, Rossi JG, et al. Automated flow cytometric MRD assessment in childhood acute B-lymphoblastic leukemia using supervised machine learning. *Cytometry A.* 2019;95(9):966–75.
59. Nissim N, Dudaie M, Barnea I, Shaked NT. Real-time stain-free classification of cancer cells and blood cells using interferometric phase microscopy and machine learning. *Cytometry A.* 2020;99:511–23.
60. Zhang ML, Guo AX, Kadauke S, Dighe AS, Baron JM, Sohani AR. Machine learning models improve the diagnostic yield of peripheral blood flow cytometry. *Am J Clin Pathol.* 2020;153(2):235–42.
61. Ko H, Chung H, Kang WS, Park C, Kim DW, Kim SE, et al. An artificial intelligence model to predict the mortality of COVID-19 patients at hospital admission time using routine blood samples: development and validation of an ensemble model. *J Med Internet Res.* 2020;22(12):e25442.

62. Van Gassen S, Vens C, Dhaene T, Lambrecht BN, Saeys Y. FloReMi: flow density survival regression using minimal feature redundancy. *Cytometry A*. 2016;89(1):22–9.
63. Yu WH, Su D, Torabi J, Fennessey CM, Shiakolas A, Lynch R, et al. Predicting the broadly neutralizing antibody susceptibility of the HIV reservoir. *JCI Insight*. 2019;4(17):e130153.
64. Dharap P, Rimbault S. Performance evaluation of machine learning-based infectious screening flags on the HORIBA Medical Yumizen H550 Haematology Analyzer for vivax malaria and dengue fever. *Malar J*. 2020;19(1):429. <https://doi.org/10.1186/s12936-020-03502-3>. PMID: 33228680; PMCID: PMC7684750
65. Frimpong A, Kusi KA, Tornyigah B, Ofori MF, Ndifon W. Characterization of T cell activation and regulation in children with asymptomatic Plasmodium falciparum infection. *Malar J*. 2018;17(1):263.
66. Jadhav S, Oswal J. Automated cellular indices to identify dengue and malaria and distinguish them from other febrile illnesses. *Int J Curr Adv Res*. 2018;7: 12176–90.
67. Dharap P, Rimbault S, Arnauvelhe S, Dray G, Janaqi S, Plantie M, et al. Validation of HORIBA Medical Pentra 80XL/XLR and MicrosemiCRP malaria flag performance derived from algorithmic data-mining techniques. *Int J Lab Hematol*. 2017;39 (suppl 2):33.
68. Briggs C, Da Costa A, Freeman L, Aucamp I, Ngubeni B, Machin SJ. Development of an automated malaria discriminant factor using VCS technology. *Am J Clin Pathol*. 2006;126:691–8.
69. Campos FMF, Repoles LC, de Araújo FF, Peruhype-Magalhães V, Xavier MAP, Sabino EC, et al. Usefulness of FC-TRIPLEX Chagas/Leish IgG1 as confirmatory assay for non-negative results in blood bank screening of Chagas disease. *J Immunol Methods*. 2018;455:34–40.
70. Doan M, Sebastian JA, Caicedo JC, Siegert S, Roch A, Turner TR, et al. Objective assessment of stored blood quality by deep learning. *Proc Natl Acad Sci U S A*. 2020;117(35):21381–90.
71. Wu YY, Huang TC, Ye RH, Fang WH, Lai SW, Chang PY, et al. A hematologist-level deep learning algorithm (BMSNet) for assessing the morphologies of single nuclear balls in bone marrow smears: algorithm development. *JMIR Med Inform*. 2020;8(4):e15963.
72. Shung DL, Au B, Taylor RA, Tay JK, Laursen SB, Stanley AJ, et al. Validation of a machine learning model that outperforms clinical risk scoring systems for upper gastrointestinal bleeding. *Gastroenterology*. 2020;158(1):160–7.
73. Venkatesh KK, Strauss RA, Grotegut CA, Heine RP, Chescheir NC, Stringer JSA, et al. Machine learning and statistical models to predict postpartum hemorrhage. *Obstet Gynecol*. 2020;135(4):935–44.
74. Lee SJ, Klein J, Haagenson M, Baxter-Lowe LA, Confer DL, Eapen M, Fernandez-Vina M, Flomenberg N, Horowitz M, Hurley CK, Noreen H, Oudshoorn M, Petersdorf E, Setterholm M, Spellman S, Weisdorf D, Williams TM, Anasetti C. High-resolution donor-recipient HLA matching contributes to the success of unrelated donor marrow transplantation. *Blood*. 2007;110:4576–83.
75. Marino SR, Lee SM, Binkowski TA, Wang T, Haagenson M, Wang HL, Maiers M, Spellman S, van Besien K, Lee SJ, Garrison T, Artz A. Identification of high-risk amino-acid substitutions in hematopoietic cell transplantation: a challenging task. *Bone Marrow Transplant*. 2016;51:1342–9.
76. Tabbara I, Zimmerman K, Morgan C, Nahleh Z. Allogeneic hematopoietic stem cell transplantation: complications and results. *Arch Intern Med*. 2002;162: 1558–66.
77. Stafford IS, Kellermann M, Mossotto E, Beattie RM, MacArthur BD, Ennis S. A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. *NPJ Digit Med*. 2020;3:30. <https://doi.org/10.1038/s41746-020-0229-3>. PMID: 32195365; PMCID: PMC7062883
78. Bransky A, Larsson A, Aardal E, Ben-Yosef Y, Christenson RH. A novel approach to hematology testing at the point of care. *J Appl Lab Med*. 2020;6:532.
79. Ben-Yosef Y, Marom B, Hirshberg G, et al. The HemoScreen, a novel haematology analyser for the point of care. *J Clin Pathol*. 2016;69:720–5.
80. Yokoyama K. Artificial intelligence-guided precision medicine in hematological disorders. *Rinsho Ketsueki*. 2020;61(5):554–63.
81. Manfredini AF, Malagoni AM, Litmanen H, Zhukovskaja L, Jeannier P, Dal Follo D, et al. Performance and blood monitoring in sports: the artificial intelligence evoking target testing in antidoping (AR.I.E.T.T.A.) project. *J Sports Med Phys Fitness*. 2011;51 (1):153–9.



Artificial Intelligence in Medicine in Anemia

104

Adam E. Gaweda and Michael E. Brier

Contents

Introduction	1442
Artificial Intelligence Tools for Anemia Management	1442
Pathophysiology and Treatment of Anemia in Chronic Kidney Disease	1442
Application of Artificial Intelligence to Diagnosis and Management of Anemia	1443
Expert Systems	1443
Moving Past Just Imitation	1445
Artificial Neural Networks	1445
Reinforcement Learning	1446
Fuzzy Systems	1447
Conclusion	1450
References	1450

Abstract

Anemia is a common comorbidity of chronic kidney disease. The abundance of clinical data readily available in electronic medical records facilitates the development of Artificial Intelligence techniques to support the human expert in diagnosis and treatment of this condition.

This chapter reviews applications of various Artificial Intelligence techniques through time to diagnose and treat anemia in patients with declining kidney function. First, Artificial Intelligence tools in this area were reported in the 1980s and focused on building computerized Expert System capable of mimicking human expert in the diagnosis of anemia. With the technological development and the ongoing data growth, more recent applications of Artificial Intelligence to the management of anemia moved from just imitating the human expert toward actively supporting them by elucidating relevant information hidden in the data.

A. E. Gaweda (✉)

Division of Nephrology and Hypertension, University of Louisville, Louisville, KY, USA

e-mail: adam.gaweda@louisville.edu

M. E. Brier

Division of Nephrology and Hypertension, University of Louisville, Louisville, KY, USA

Robley Rex Veterans Administration Medical Center,
Louisville, KY, USA

e-mail: michael.brier@louisville.edu

© Springer Nature Switzerland AG 2022

N. Lidströmer, H. Ashrafiyan (eds.), *Artificial Intelligence in Medicine*,

https://doi.org/10.1007/978-3-030-64573-1_183

1441

Keywords

Machine learning · Anemia · Erythropoietin · Nephrology · Artificial Intelligence · Model

predictive control · Reinforcement Learning · Expert System · Fuzzy systems

Introduction

The term Artificial Intelligence (AI) and its application have evolved over time as computing capabilities have increased. In 1950, Alan Turing posed the following question: “can machines think?” [1]. He proposed an experiment called “The Imitation Game” where an interviewer will question two individuals to determine which is male and which is female. In the game, one of the test subjects is replaced with a computer, and the goal is to determine how often the interviewer wrongly guesses the sex of the subjects. One might conclude that the machine is “thinking” like the subject that was replaced when the proportions mimic those of human subjects. At the end of this work, he states “We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with?”

In this case, the management of anemia is a perfect test bed in which to apply an AI approach in our own version of “The Imitation Game.” We would hope that we could not only develop a computer that is capable of mimicking the actions of a physician well versed in the management of anemia of chronic kidney disease (CKD) but also exceed the performance of a human expert. Anemia management in patients receiving hemodialysis is a good choice to demonstrate this research due to the nature of the dialysis process and a tightly controlled environment in which it occurs. The inputs to the process (drug dose, hemoglobin, time, etc.) are well captured in this patient population due to reporting and billing requirements and the very nature of the care provided for these patients. The outcome of interest is also well defined as maintaining hemoglobin concentration without exceeding specific upper and lower levels as well as minimizing the number of transfusion events. Errors related to adherence to therapy are minimized since all treatments are captured. The frequency of observation of the process allows us to observe the dynamics of response and provide

timely dose corrections. Although we focus on the use of AI in the management of anemia, it is also used for prediction of the reason for anemia as will be shown.

Artificial Intelligence Tools for Anemia Management

Pathophysiology and Treatment of Anemia in Chronic Kidney Disease

The oxygen-carrying ability of the blood stimulates the release of erythropoietin, primarily by the kidney but also other organs [2]. When patients develop kidney disease and end up on dialysis, the production of and the body’s response to erythropoietin become significantly impaired. Recurring loss of blood during dialysis sessions, reduced red blood cell lifespan, alteration in erythropoietin production and sensitivity, decrease of iron availability, and systemic inflammation lead to anemia in this patient population. In 1989, recombinant human erythropoietin was approved as the first erythropoiesis-stimulating agent (ESA) for treatment of anemia in the United States. Prior to that time, red blood cell transfusions were the primary mode of treatment.

To treat anemia of CKD, human experts attempt to mimic the role that kidneys play in this process by assessing the hemoglobin and iron levels and administering ESA and supplementing iron as needed. Physiologically, this process is precisely controlled at a relatively rapid timescale. Clinically, the precision is limited by the dose availability from the drug manufacturer and longer-time intervals between subsequent dose adjustments. Due to the complex relationship between the ESA dose, red blood cell lifespan, iron availability, and ESA response, as well as the logistic limitations listed above, we contended that an AI-supported approach to dose monitoring and outcome prediction would enable improvements in the treatment of anemia of CKD.

In most dialysis-dependent CKD patients, hemodialysis treatment is performed three times per week. Many therapeutic agents, including those used for anemia treatment, are administered

at that time. Laboratory measurements of hemoglobin and other routinely measured blood parameters are typically performed on a monthly or quarterly basis. Some of these measurements may be performed as frequently as once a week. As a result, data sets exist with very reliable data that can be used for developing AI approaches for monitoring and therapy. Laboratory measurements are usually downloaded into electronic medical record (EMR) database within 48 h of their collection, and ESA and iron doses are entered into the electronic medical record as they are given. Dose recommendations are typically governed by the information gathered by the manufacturer during development of the drug. Drug studies, performed in different patient cohorts, determine both the pharmacokinetic and pharmacodynamic properties of the drug. Knowledge gained during these studies is incorporated into the drug package insert issued by the Food and Drug Administration (FDA). Specifically, the package insert for anemia medications recommends using the lowest ESA dose required to decrease the need for red blood cell transfusions. Dose adjustments should consider the rate of change of the hemoglobin concentration, individual patient responsiveness to the drug, and hemoglobin variability. Doses should not be increased more often than once every 4 weeks, and the rate of hemoglobin rise should not exceed 1 g/dL in a 2-week period. Despite the existence of accurate models of the drug effect in patients, these models are not used in clinical practice in this specific instance and for all drugs in general. Rather, recommendations are given based on the population as a whole and not for an individual. Because of this, dialysis providers often establish their own algorithms guiding initial drug dose selection and the subsequent dose adjustments. In many cases, these algorithms are a form of an Expert System, where a physician expert in anemia management defines rules by which drug dosing performed.

In the United States, approximately two-thirds of dialysis patients receive treatment in one of two large dialysis organizations. The remaining one-third of dialysis patients receive treatment in medium and small dialysis organizations. This can result in different algorithms for the

administration of erythropoietic drugs. While this process may be relatively standardized in the large dialysis facilities, opinions become more varied as we look at smaller facilities. To this end, in partnership with End-Stage Kidney Disease Networks 9 and 10 based in Indianapolis, IN, the performance of six facility-supplied anemia management algorithms was tested in silico with the results shown in Fig. 1 (unpublished data). These results show that individual facility interpretation of the drug dosing guidelines provided in the FDA package insert for ESA can vary significantly and illustrate the nature of the problem.

Application of Artificial Intelligence to Diagnosis and Management of Anemia

Several areas of what one would consider application of AI to anemia management have been explored and will be discussed in the following sections. In addition, other non-AI approaches have been proposed that incorporate pharmacodynamic models for dose recommendation [3, 4]. These alternative approaches are outside the scope of this chapter.

Expert Systems

Expert Systems have been widely used in medicine and are the first form of AI that was applied to anemia treatment. The first reported use of an AI-based Expert System in this area occurred in 1985 with the development of a knowledge-based consultation program called ANEMIA [5–7]. This Expert System was developed to match the cognitive performance of an expert hematologist in diagnosing 65 specific variations of anemia from patient's laboratory results. In validation, using a process akin to Turing's "Imitation Game," the diagnostic performance of ANEMIA was rated as acceptable by expert hematologists in close to 90% of the cases. Similar successful examples of early AI-based Expert Systems for differential diagnosis of anemia are presented in [8–13]. Of

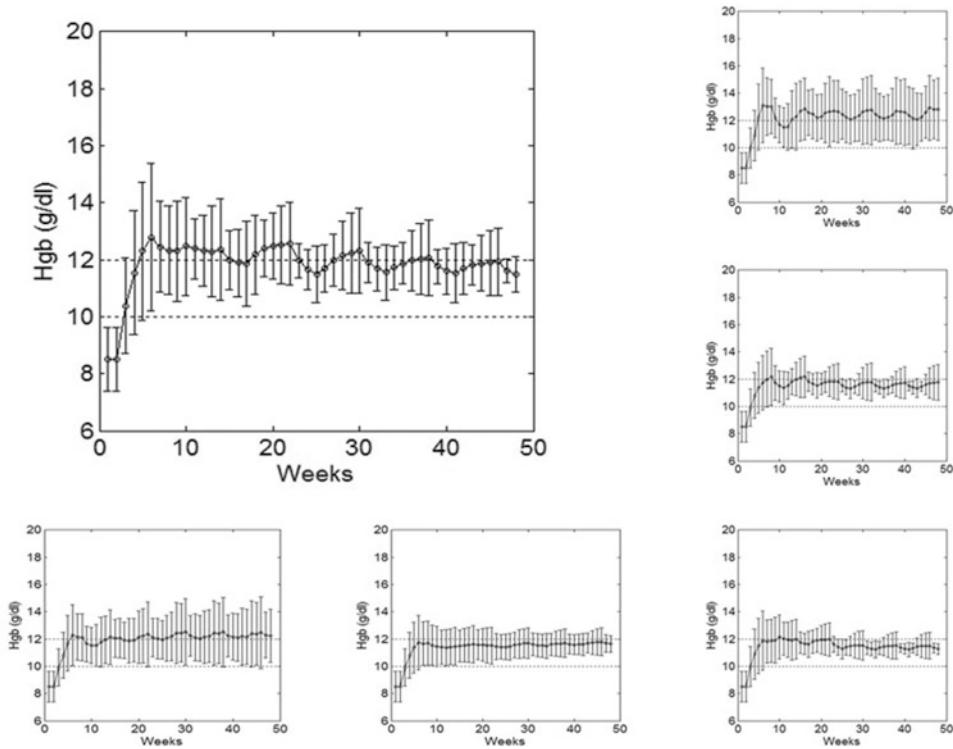


Fig. 1 Simulation of six different anemia management protocols designed by facilities when the hemoglobin target was 10.0–12.0 mg/dL

note, the hybrid Expert System proposed in [14] combined the knowledge-driven rule-based and a data-driven Artificial Neural Network (ANN) approach to AI. The latter will be addressed in greater detail in the following sections.

The first use of an Expert System to provide drug dosing information was tested in a randomized controlled clinical trial [15]. This system was developed by several investigators proficient in the management of CKD anemia and consisted of the following: assessment of hemoglobin concentration every 4 weeks, division of the obtained hemoglobin into three hemoglobin ranges, and for hemoglobin in the lowest range determining if iron stores are adequate, then recommending a step-up in dose, stepping down in dose, dose holding, or administering intravenous iron. This approach represented the computer implementation of a traditional paper algorithm like those used to generate the data shown in Fig. 1. It would meet Turing's definition of "intelligence"

where the recommendations could not be discerned from those made by a practitioner experienced in the treatment of anemia. However, due to the subjective nature of the expert's knowledge, it cannot be guaranteed that the dose recommendations generated by such system are optimal with respect to the published FDA guidelines or best practice.

A similar approach has been proposed by investigators who developed a clinical decision support system for the management of anemia in CKD patients [16]. This system utilizes the concept of threshold values for ESA dose changes in individual patients [17]. This work has resulted in the development of a commercial product called AMIE (Media Innovations, University of Leeds, Leeds, UK). This tool provides not only ESA dose recommendations but also dose for intravenous iron. Testing of the program in a 24-month prospective study in a cohort of hemodialysis patients resulted in a 50% decrease in ESA dose as well as

better iron status management. The decrease in ESA dose administered while maintaining hemoglobin concentrations within the target range is a result that will be seen in almost every other application of AI to support anemia management.

The dose recommendation Expert Systems described above only looked at cross-sectional data and did not benefit from longitudinal data on ESA dose administration and hemoglobin measurements readily available in today's electronic medical records. Dose recommendations were based solely on a single observation. Likewise, these Expert Systems did not address the issue of hemoglobin variability and inter-patient variability in ESA response. Therefore, the impact of these Expert Systems on decreasing individual hemoglobin variability remains unknown [18].

Moving Past Just Imitation

To better leverage the information collected and stored in an electronic medical record, AI approaches have been applied to the anemia of CKD. These approaches attempt to better accomplish the dosing recommendations provided in the FDA-approved package insert. The pharmacologic management of the disease is typically performed in a trial-and-error fashion by physicians. This management can be augmented by therapeutic drug monitoring where the concentration of the administered drug in the blood, or in the case of anemia management of the hemoglobin concentration, is measured, and a new dose is recommended as a function of that measured concentration. Physiologic data is complicated by error and variability. Error can occur in the data in the form of bias in the measurement of the hemoglobin concentration as well as adherence to drug therapy. Fortunately, in hemodialysis, adherence to drug therapy is known. However, there can be considerable variability in the measurement of hemoglobin concentration in dialysis patients due to assay variability, timing of the blood sample during the dialysis session, and dilution of the hemoglobin in the blood due to patient's variable fluid intake prior to the measurement. Physiologic variability can occur due to

differences in the metabolism and elimination of the administered drug as well as differences at the level of the erythropoietin receptor that is responsible for erythropoiesis. A more robust method may be needed to account for all these shifting variables in the prediction problem, and we can see the need to move past basic expert systems to provide optimal therapy.

Artificial Neural Networks

The most frequently applied AI tool in the management of CKD anemia is the Artificial Neural Network (ANN). An ANN represents an empirical, strictly data-driven AI approach. In the context of anemia management, ANNs use large quantities of data to look for associations between patient factors, dose, and response to predict future hemoglobin levels and/or recommend new erythropoietin dose. This approach is most useful when large amounts of data are available, and the dose-response relationship and its predictors are complex and may not be easily understood. In the United States approximately 600,000 patients receive in-center hemodialysis, and most of them receive some form of erythropoiesis-stimulating agent (ESA). Therefore, sufficient data should be available to successfully train an ANN for hemoglobin prediction. Following the work published in the mid-1980s with an Expert System, this was the next form of AI applied to this problem.

Published simultaneously in 2003, two groups researching the use of ANNs presented their results [19, 20]. This initial work focused on developing a pharmacodynamic model representing hemoglobin response to erythropoietin. The work published by Gaweda et al. [19] compared two common types of ANNs, a multilayer perceptron (MLP) network and a radial basis function (RBF) network to a linear auto-regressive model (ARX), and demonstrated the superiority of the MLP network. Martin-Guerrero et al. at the University of Valencia [20, 21] performed a similar study comparing an MLP network and a support vector machine (SVM) to an auto-regressive model. The authors concluded that the results of predicting future hemoglobin were very similar between the MLP

and SVM methods. As presented by these two papers, the very first step in the development of a drug dosing tool is a dose-response model which can be used to in turn calculate ESA dose.

The goal of developing a model for prediction of hemoglobin concentration (pharmacodynamic model) is that one can use that model to advise a medical expert as to the most appropriate dose of erythropoietin needed to achieve target hemoglobin. One must then couple the pharmacodynamic model with some other tool to predict dose or back solve the relationship for the dose needed to see the appropriate change in hemoglobin concentration. One way this can be accomplished is using closed-loop control. Brier et al. report on the use of an ANN-based model predictive control (MPC) to make ESA dosing recommendations [22]. The work represents the use of the gold standard for demonstrating the utility of an intervention, the randomized, double-blind, controlled trial. The study consisted of 60 patients randomly assigned to a control and treatment group, where the control group received treatment guided by the standard of care. Dose recommendations were provided by the ANN-MPC algorithm and a group of nurse practitioners for all study subjects. An unblinded physician entered the appropriate dose for each subject into the electronic medical record based on the subject's group assignment. Subjects were followed for 8 months, and the proportions of measured hemoglobin within the target range of 11–12 g/dL, less than 9 g/dL, and above 13 g/dL were measured. The use of the ANN-MPC resulted in a 50% decrease in outlier hemoglobin values and a decrease in the mean absolute difference between the achieved hemoglobin and 11.5 g/dL. Surprisingly, subjects dosed with the ANN-MPC algorithm had a slight increase in ESA utilization compared to the control group unlike the trial presented above in the expert system section.

A randomized controlled clinical trial is usually statistically powered to demonstrate the impact of a treatment on the primary outcome variable. However, an approach which may be successful in a small, controlled environment may potentially lack external validity. Barbieri et al. [23, 24] used a similar approach as the one

discussed above and developed an ANN-based algorithm using anemia management data from 4135 subjects in a large dialysis organization. Owing to the availability of a large data set, they expanded the inputs used for the prediction to include history of hemoglobin levels and ESA dose, patient characteristics, dialysis prescription, iron parameters, as well as markers of nutrition, infection, and inflammation. The ANN algorithm predicted hemoglobin achieved at 3 months. The prediction made by the ANN model was used to choose the most desirable dose to maintain hemoglobin within the target range. The ANN-driven anemia management protocol that was developed was prospectively validated in three dialysis clinics. The investigators reported a 6% absolute improvement in hemoglobin maintenance within the 10–12 g/dL target range, a decrease in individual hemoglobin variability, as well as 25% dose reduction compared to standard-of-care anemia management used prior to the study [24].

Reinforcement Learning

Drug dosing can be viewed as a form of a trial-and-error learning process within a feedback loop. After the initial drug dose is administered, the patient is observed for specific physiologic responses or adverse events. Subsequently, the practitioner adjusts the dose following the observed state of the patient. If toxicity occurs, the dose amount is decreased. If an inadequate response is observed, the dose is increased. The trial-and-error process continues until a desired response is achieved. Reinforcement Learning (RL) is an AI methodology rooted in the psychological theory of learning that mimics the human trial-and-error learning process [25]. One of the advantages of using this methodology is that some of the RL algorithms could generate decision rules that explain the actions taken. This feature provides an additional source of validation for the expert user. To study the feasibility of the RL approach in anemia management, several investigators tested generating ESA dose recommendations using an RL technique called Q-learning [26–30].

Gaweda et al. [26, 30] first used the RL approach to simulate anemia management in 200 virtual patients including a group of patients who responded normally to erythropoietin as well as a group with an impaired response. Simulation results were compared to a standard anemia management approach showing comparable performance. In the poor responder group, the standard approach typically resulted in the attainment of the target hemoglobin more rapidly. This phenomenon was attributed to the learning occurring at the beginning of the RL-driven treatment and a more conservative initial dose choice compared to the standard approach. To improve the speed and the efficiency of the Q-learning algorithm applied to anemia management, the same investigators proposed a method to introduce the expert knowledge into the learning updates [29].

The University of Valencia group also investigated the use of RL for the individualization of recombinant ESA dosage [27]. The approach used is like that of Gaweda et al. listed above and compares an RL approach to a standard anemia management protocol. These investigators compared standard interactive Q-learning approach to fitted Q-iteration, which is an off-line data-driven modification of the algorithm. In their experience, standard Q-learning resulted in a greater dispersion of hemoglobin values around the target range of 11–12 g/dL, compared to fitted Q-iteration and the standard of care. Several outlier hemoglobins were observed for all three methodologies. In the simulations, the fitted Q-iteration approach resulted in a 27.6% increase in the proportions of hemoglobin values in the target range compared to the standard protocol and a decrease in drug use of 5.13%.

Reinforcement Learning approaches have been investigated in several simulations. Depending on the RL method applied and the learning setup, different levels of success have been achieved. The advantage of the RL approach in the generation of rules that can be explicitly communicated to a practitioner makes this approach intriguing. At this point, sufficient work has not been performed to estimate the impact that the RL approach would have on patient outcomes in anemia management.

Fuzzy Systems

Returning to our discussion on the drug development process and the data and studies used for the registration of a new drug, it is primarily dependent on the manufacture to demonstrate safety and efficacy. Additionally, they are required to provide drug dosing information which usually is population-based with some exceptions for sub-populations. In the case of ESA, in the section on dosage and administration, there are specific recommendations for chronic renal failure, zidovudine-treated HIV-infected, cancer, and surgery patients. The recommendations acknowledge that individual patients are different for several reasons and that “doses must be individualized to ensure that hemoglobin is maintained at the appropriate level for each patient” (EPOGEN FDA product label). Fuzzy set theory may be helpful in identifying patients into groups of similar patients and that membership is not binary (or otherwise).

Unlike a binary factor like sex as defined by XX or XY which is easily ascertained in most cases, drug responsiveness is a blend of genetic and environmental factors. In the case of Mendelian genetics where each inherited trait is defined by a gene pair, this can result in three cases: AA, AB, or BB. Using a genotypic approach, it is easy to assign individuals to specific groups. This is not the case for a phenotypic approach especially when environmental factors come into play. This is where a fuzzy set approach can be useful. As shown in Fig. 2, using a fuzzy clustering approach, patient can be defined as a member of two or more groups with the centroid observed as the peak in the distribution for each cluster. The number of clusters is often determined empirically.

Once again, the goal of information provided to the practitioner in the drug package insert is based on the population than the individual. The clinic is interested in providing dose recommendations for the individual patient. The process by which a dose of drug is determined for an individual patient is typically guided by information contained in the FDA-required drug package insert and supplemented which patient-specific information and the clinician’s best judgment.

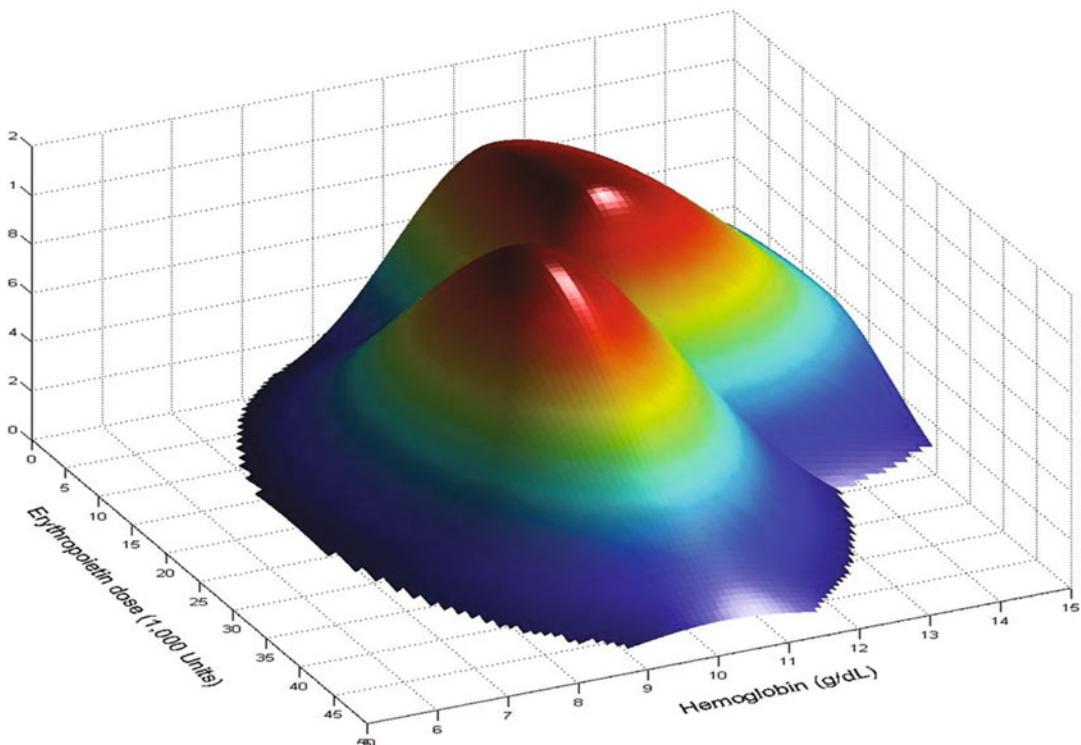


Fig. 2 Fuzzy membership functions representing patient classification as a “good responder” (low erythropoietin dose, high hemoglobin) and “poor responder” (high erythropoietin dose, low hemoglobin)

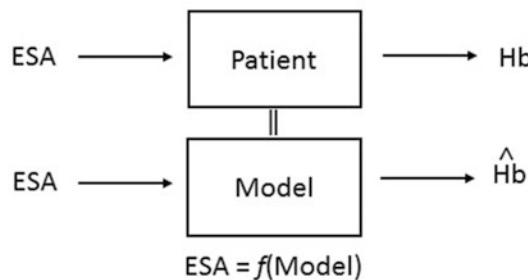
However, these approaches have limitations due to the data used to develop them. An AI-based closed-loop control can be used to account for imperfections in the model of the patient used to predict hemoglobin response. In an open-loop approach, future ESA dose is determined either by a written set of instructions or presenting different ESA doses to a patient model. In closed-loop control, the control system recommends an ESA dose which is given to the patient resulting in a hemoglobin value. This value is then compared to the hemoglobin target, and an error signal is generated. This error signal is fed back to optimize the subsequent ESA dose (Fig. 3).

The closed-loop control was used in the only randomized controlled clinical trials of AI in the management of anemia of CKD [22, 31] where the concepts of a fuzzy classification (Fig. 2) were combined with a model predictive control (MPC). In these two publications, MPC was applied in one instance where the patient model was an ANN and in the second instance where

the patient model was a physiologic model of erythropoiesis. Comparisons were made to a standard anemia management protocol developed by the dialysis facility medical director. Both studies were powered to detect a difference in the hemoglobin variability within and between subjects. As such, the size of the studies is limited to 60 patients. The ANN-MPC approach used only Hb and ESA dose inputs for the prediction of the next hemoglobin value. Future hemoglobin values were predicted over a 3-month horizon. Subjects were followed for 8 months. The Artificial Neural Network-based model predictive control group resulted in decreased hemoglobin variability in fewer observations below 9 and greater than 13 mg/dL. Total ESA dose was increased in the ANN group.

A second randomized controlled clinical trial of MPC-guided anemia management was performed using an AI tool Smart Anemia Manager (SAM). Subjects were followed for 12 months following randomization, and the

Open Loop



Closed Loop

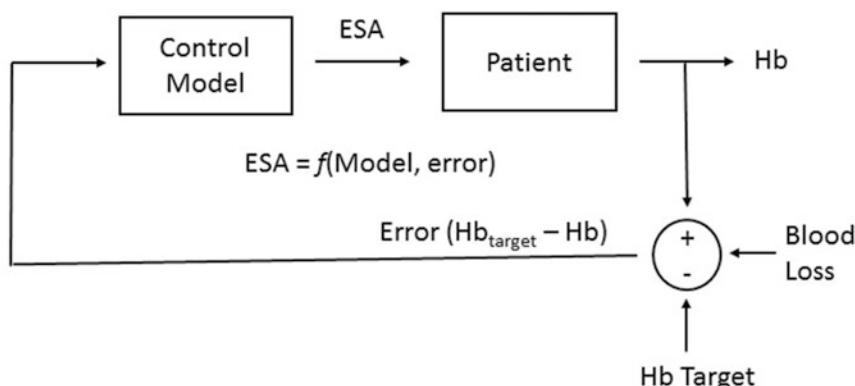


Fig. 3 Comparison between open-loop and closed-loop approaches to anemia management

control group was once again an anemia management protocol developed by the facility medical director. The study was designed to decrease within subject hemoglobin variability. As expected, the hemoglobin variability decreased (from 1.5 to 1.2 g/dL). This decrease led to an increase in the number of hemoglobin observations within the target range of 10–12 mg/dL from 61.9% in the control group to 72.5% in the treatment group. Median ESA and iron dose were not different between the two groups. Secondary analysis of the data observed a 50% reduction in transfusion rate in the treatment group.

Long-term follow-up of the SAM tool was subsequently performed. SAM was adopted as the standard-of-care anemia management protocol for the dialysis facility in which it was developed. Two follow-up periods were defined: (1) those patients on dialysis and treated with a facility-developed anemia management protocol

from September 2009 through August 2010 were compared to SAM dosing from August 2012 through July 2013 and (2) long-term follow-up of all in-center hemodialysis patients from October 2013 through March 2017. The ESA product used in both follow-up periods was epoetin alfa. In the case-control analysis, each patient served as their own control, and a total of 56 patients were analyzed. SAM-guided ESA dosing resulted in a 7% increase in hemoglobin concentrations within the target range of 10–12 g/dL. SAM-guided dosing also resulted in fewer hemoglobin measurements less than 10 g/dL by about 10% and greater than 12 g/dL by about 17%. ESA dose decreased by 62% and iron dose by 63%. The long-term follow-up in 233 patients receiving in-center hemodialysis was performed for 42 months and terminated when the dialysis facility switched to a different ESA product. The percentage of hemoglobin observations within the

target range was maintained between 68% and 77% for the fourth quarter of 2013 through the first quarter of 2017. ESA dose received continued to remain below mean dose established during the control period of 7000 units/week ranging from 2315 to 4176 units/week. Confidence in this technology is further demonstrated by the implementation of the SAM software as a clinical decision support tool and use in a medium-sized dialysis provider using a different, long-acting ESA product, darbepoetin. A total of 3354 patients were available for dosing using the SAM tool. 2946 patients had 3 months of prior dosing data for comparison along with 12 months of follow-up on SAM. These results were published at the 2019 Meeting of the American Society of Nephrology. Mean hemoglobin values in the control for 3 months were 10.3 ± 1.3 g/dL and ranged from 10.3 ± 1.3 g/dL in the first quarter of use to 10.5 ± 1.3 g/dL in the fourth quarter of use. ESA dose fell from a median monthly dose of 87 µg in the control period and ranged from 67 µg to 43 µg in quarters 1 and 4, respectively.

Iron is equally important to erythropoietin in the production of red blood cells. One AI-based approach to diagnose iron deficiency anemia has been presented in this area using an adaptive neuro-fuzzy inference system (ANFIS) [32]. Results obtained by the ANFIS were compared to those produced by an ANN, logistic, and linear regression using receiver operating characteristic (ROC) curve. Diagnostic accuracy improved over logistic regression ($AUC = 0.691$) with both, the ANN ($AUC = 0.982$) and the ANFIS ($AUC = 0.951$). In a total of 54 observations in a test set, the ANN only had 1 false negative and 1 false positive as compared to the logistic regression analysis with 18 false negatives and 2 false positives. This work demonstrates another successful use of an AI approach to diagnose anemia.

Conclusion

Artificial Intelligence tools have been successfully applied to the diagnosis and treatment of anemia. AI has been prominently utilized in the

medical literature for classification. The more novel use is in the treatment of anemia of CKD. The initial uses of AI for treatment were in the form of expert systems which had the advantage of uniformly applying rules to ESA dosing and removing individual practitioner variability from the equation. As individuals acquainted with AI approaches became aware of the unique prediction environment provided by in-center hemodialysis, the use of AI approaches expanded. The most common tool applied to this prediction problem is the Artificial Neural Networks. Other approaches including Reinforcement Learning and Fuzzy Systems have been studied as well. There is now sufficient literature evidence to support the use of AI tools as an addition to standard-of-care approaches to anemia management. These AI tools may not only improve the patient outcomes and the cost-effectiveness of treatment but also reduce and streamline the workload, allowing the medical personnel to better manage their effort.

References

1. Turing AM. Computing machinery and intelligence. *Mind*. 1950;49:433–60.
2. Singh AK. Erythropoiesis: the roles of erythropoietin and iron. In: Singh AK, Williams GH, editors. *Textbook of nephro-endocrinology*. 2nd ed. London: Elsevier; 2018.
3. Uehlinger DE, Gotch FA, Sheiner LB. A pharmacodynamic model of erythropoietin therapy for uremic anemia. *Clin Pharmacol Ther*. 1992;51(1):76–89.
4. McCarthy JT, Hocum CL, Albright RC, Rogers J, Gallaher EJ, Steensma DP, et al. Biomedical system dynamics to improve anemia control with darbepoetin alfa in long-term hemodialysis patients. *Mayo Clin Proc*. 2014;89(1):87–94.
5. Barosi G, Berzuini A, Quaglini S, Stefanelli M. Anemia: an artificial intelligence project. *Haematologica*. 1985;70(4):358–62.
6. Quaglini S, Stefanelli M, Barosi G, Berzuini A. ANEMIA: an expert consultation system. *Comput Biomed Res*. 1986;19(1):13–27.
7. Quaglini S, Stefanelli M, Barosi G, Berzuini A. A performance evaluation of the expert system ANEMIA. *Comput Biomed Res*. 1988;21(4):307–23.
8. Blomberg DJ, Guth JL, Fattu JM, Patrick EA. Evaluation of a new classification system for anemias using Consult Learning System. *Comput Methods Prog Biomed*. 1986;22(1):119–25.

9. Blomberg DJ, Ladley JL, Fattu JM, Patrick EA. The use of an expert system in the clinical laboratory as an aid in the diagnosis of anemia. *Am J Clin Pathol.* 1987;87(5):608–13.
10. Houwen B. The use of inference strategies in the differential diagnosis of microcytic anemia. *Blood Cells.* 1989;15(3):509–27; discussion 27–32.
11. Imbert M, Priolet G, Dadi W, Sultan C. An expert system applied to the diagnosis of anemia with special reference to myelodysplastic syndromes. *Blood Cells.* 1989;15(3):563–70; discussion 70–1.
12. O'Connor ML, McKinney T. The diagnosis of microcytic anemia by a rule-based expert system using VP-Expert. *Arch Pathol Lab Med.* 1989;113(9):985–8.
13. Sultan C, Imbert M, Priolet G. Decision-making system (DMS) applied to hematology. Diagnosis of 180 cases of anemia secondary to a variety of hematologic disorders. *Hematol Pathol.* 1988;2(4):221–8.
14. Birndorf NI, Pentecost JO, Coakley JR, Spackman KA. An expert system to diagnose anemia and report results directly on hematology forms. *Comput Biomed Res.* 1996;29(1):16–26.
15. Brimble KS, Rabbat CG, McKenna P, Lambert K, Carlisle EJ. Protocolized anemia management with erythropoietin in hemodialysis patients: a randomized controlled trial. *J Am Soc Nephrol.* 2003;14(10):2654–61.
16. Will EJ, Richardson D, Tolman C, Bartlett C. Development and exploitation of a clinical decision support system for the management of renal anaemia. *Nephrol Dial Transplant.* 2007;22(Suppl 4):iv31–iv6.
17. Richardson D, Bartlett C, Will EJ. Intervention thresholds and ceilings can determine the haemoglobin outcome distribution in a haemodialysis population. *Nephrol Dial Transplant.* 2000;15(12):2007–13.
18. Berns JS, Fishbane S. Hemoglobin variability: random fluctuation, epiphenomenon, or phenomenon? *Semin Dial.* 2006;19(3):257–9.
19. Gaweda AE, Jacobs AA, Brier ME, Zurada JM. Pharmacodynamic population analysis in chronic renal failure using artificial neural networks—a comparative study. *Neural Netw.* 2003;16(5–6):841–5.
20. Martin-Guerrero JD, Olivas ES, Valls GC, Serrano-Lopez AJ, Perez-Ruixo JJ, Torres NV. Use of neural networks for dosage individualisation of erythropoietin in patients with secondary anemia to chronic renal failure. *Comput Biol Med.* 2003;33(4):361–73.
21. Martin-Guerrero JD, Camps-Valls G, Soria-Olivas E, Serrano-Lopez AJ, Perez-Ruixo JJ, Jimenez-Torres NV. Dosage individualization of erythropoietin using a profile-dependent support vector regression. *IEEE Trans Biomed Eng.* 2003;50(10):1136–42.
22. Brier ME, Gaweda AE, Dailey A, Aronoff GR, Jacobs AA. Randomized trial of model predictive control for improved anemia management. *Clin J Am Soc Nephrol.* 2010;5(5):814–20.
23. Barbieri C, Bolzoni E, Mari F, Cattinelli I, Bellocchio F, Martin JD, et al. Performance of a predictive model for long-term hemoglobin response to darbepoetin and iron administration in a large cohort of hemodialysis patients. *PLoS One.* 2016;11(3):e0148938.
24. Barbieri C, Molina M, Ponce P, Tothova M, Cattinelli I, Ion Titapiccolo J, et al. An international observational study suggests that artificial intelligence for clinical decision support optimizes anemia management in hemodialysis patients. *Kidney Int.* 2016;90(2):422–9.
25. Sutton RS, Barto AG. Reinforcement learning: an introduction. 2nd ed. Cambridge, MA: The MIT Press; 2018. xxii, 526 p
26. Gaweda AE, Muezzinoglu MK, Jacobs AA, Aronoff GR, Brier ME. Model predictive control with reinforcement learning for drug delivery in renal anemia management. Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Conference. 2006;1: 5177–5180.
27. Escandell-Montero P, Chermisi M, Martinez-Martinez JM, Gomez-Sanchis J, Barbieri C, Soria-Olivas E, et al. Optimization of anemia treatment in hemodialysis patients via reinforcement learning. *Artif Intell Med.* 2014;62(1):47–60.
28. Martin-Guerrero JD, Gomez F, Soria-Olivas E, Schmidhuber J, Climente-Marti M, Jimenez-Torres NV. A reinforcement learning approach for individualizing erythropoietin dosages in hemodialysis patients. *Expert Syst Appl.* 2009;36(6):9737–42.
29. Gaweda AE, Muezzinoglu MK, Aronoff GR, Jacobs AA, Zurada JM, Brier ME. Using clinical information in goal-oriented learning. *IEEE Eng Med Biol Mag.* 2007;26(2):27–36.
30. Gaweda AE, Muezzinoglu MK, Aronoff GR, Jacobs AA, Zurada JM, Brier ME. Individualization of pharmacological anemia management using reinforcement learning. *Neural Netw.* 2005;18(5–6):826–34.
31. Gaweda AE, Aronoff GR, Jacobs AA, Rai SN, Brier ME. Individualized anemia management reduces hemoglobin variability in hemodialysis patients. *J Am Soc Nephrol.* 2014;25(1):159–66.
32. Azarkish I, Raoufy MR, Gharibzadeh S. Artificial intelligence models for predicting iron deficiency anemia and iron serum level based on accessible laboratory data. *J Med Syst.* 2012;36(3):2057–61.



Matthieu Komorowski and Alexandre Joosten

Contents

Introduction	1454
The Use of AI to Monitor Depth of Anesthesia	1455
Controlling Anesthesia Delivery with AI	1455
Perioperative Hemodynamic Optimization Assisted by AI	1456
Automation for Fluid Therapy	1457
Automation for Vasopressor Titration	1458
Automation for Inotrope Infusion	1458
Automation for Vasodilator Infusion	1458
Machine Learning for Predicting Hypotension	1458
Event and Risk Prediction with AI	1459
Other Applications of AI in Anesthesiology	1460
Technology Readiness Level of Published Applications	1460
Implications of AI for the Anesthesiologist	1461
Conclusion	1462
References	1462

Abstract

This chapter focuses on applications of artificial intelligence (AI) in anesthesiology. Anesthesiology is the field of healthcare involved with providing a state of controlled, temporary loss of sensation or awareness that is induced for medical purposes such as a surgical intervention. It may include a combination of various components including analgesia, amnesia, unconsciousness, and muscle relaxation. Anesthesiology is mostly a technical field, heavily protocolized and data-intensive (due to all the

M. Komorowski (✉)
Department of Surgery and Cancer, Imperial College London, London, UK

Intensive Care Unit, Charing Cross Hospital, Imperial College Healthcare NHS Trust, London, UK
e-mail: m.komorowski14@imperial.ac.uk

A. Joosten
Department of Anesthesiology and Intensive Care, Hôpitaux Universitaires Paris-Sud, Université Paris-Sud, Université Paris-Saclay, Hôpital De Bicêtre, Assistance Publique Hôpitaux de Paris (AP-HP), Le Kremlin-Bicêtre, France

monitoring equipment in place), which makes it the perfect environment to deploy AI tools. In this chapter, we review in detail the main applications of AI in the operating room, sorted in four main domains: (1) monitoring of the depth of anesthesia, (2) control of administration of anesthetic drugs (hypnotics, opioids, and/or muscle relaxants), (3) hemodynamic control (mostly titration of fluids and vasopressor therapy), and (4) risk prediction and prediction of events (e.g., predict surgery length or postoperative complications). In addition, we analyzed the degree of maturity of these various technologies. While many of these applications can have a large impact on quality and safety of care surrounding anesthesia, the maturity of these technologies is in general very low, and most of the applications published describe tools that have not received prospective evaluation. Only a handful of randomized trials comparing standard of care to the tandem AI doctor could be identified. Finally, we conclude with an assessment of the current practical implications of AI for practicing anesthesiologists.

Keywords

Autonomous system · Machine learning · Artificial intelligence · Robotic anesthesia · Automation · Closed-loop system · Clinical decision support system

Introduction

This chapter focuses on applications of artificial intelligence (AI) in anesthesiology. Anesthesiology is the field of healthcare involved with providing a state of controlled, temporary loss of sensation or awareness that is induced for medical purposes such as a surgical intervention [1]. It may include a combination of various components including analgesia (relief from or prevention of pain), amnesia (loss of memory), unconsciousness, and paralysis (muscle relaxation). The two main types of anesthesia that can be provided are general (a medically induced coma) or regional (techniques that numb a large

part of the body, such as an arm, a portion of a leg, or from the waist down). Anesthesiology is mostly a technical field, heavily protocolized and data-intensive (due to all the monitoring equipment in place), which makes it the perfect environment to deploy AI tools [2]. The practice of anesthesiology is of course inevitably dependent upon technology. Anesthetics were first made possible, then increasingly made safer, and now are more innovative and skillful than ever mainly due to advances in monitoring devices and delivery technology.

There are four main domains of applications of AI in the operating room:

1. *Monitoring of depth of anesthesia:* AI is used to quantify the depth of sedation and/or analgesia.
2. *Control of administration of anesthetic drugs:* The administration of anesthetic drugs is titrated by an algorithm to achieve a desired level of sedation and pain control.
3. *Hemodynamic control.* AI is used to titrate fluid and vasopressor therapy, to predict and/or control hemodynamic instability induced by anesthetic drugs, mechanical ventilation, or surgical interventions.
4. *Risk prediction and prediction of events.* Supervised machine learning models are designed to predict patient events and outcomes and generate prediction tools and scores that can be used prospectively with new incoming patients.

The first three domains of applications are building blocks toward automated anesthesia delivery systems, where the task of the anesthesiologist is emulated by machines. Figure 1 depicts an automated closed-loop (feedback control) system, designed to maintain a given target patient parameter into a specified range, without requiring human intervention. These feedback loops achieve their purpose by feeding the measured value of the parameter as well as the target range into a controller that adjusts the rate of drug administration while taking into account patient characteristics (e.g., gender, age, body weight, etc.).

In addition, we analyzed the degree of maturity of these various technologies, by reviewing all

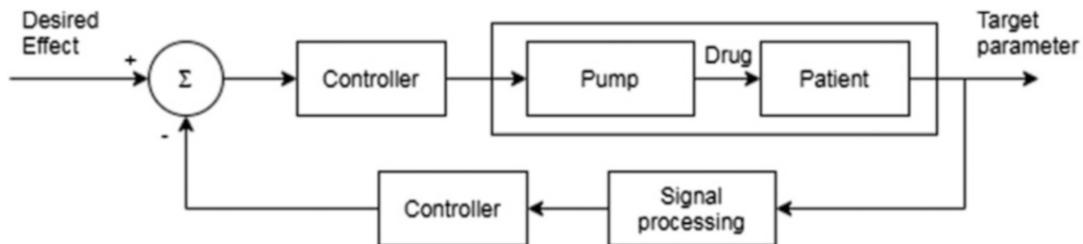


Fig. 1 Principle of a closed-loop delivery system for anesthesia. Drugs can be anesthetic agents, IV fluids, or vasopressors, and the measured target parameter can be an

EEG signal, a nociception monitor, a score of depth of hypnosis, common hemodynamic parameters, etc.

137 publications listed in a recent review of the literature on the topic [2]. While many of these applications can have a large impact on quality and safety of care surrounding anesthesia, the maturity of these technologies is in general very low, and most of the applications published in the literature describe tools that have only been developed on data collected retrospectively, with no prospective evaluation, and very few randomized trials comparing standard of care to the tandem AI doctor. Finally, we propose an assessment of the current practical implications of AI for practicing anesthesiologists.

The Use of AI to Monitor Depth of Anesthesia

The bulk of the literature in AI for anesthesiology relates to techniques trying to quantify the depth of anesthesia or analgesia, with the earliest models now about 30 years old [3, 4]. All these models belong to the realm of supervised learning and attempt to produce a value of awareness (regression models [5]) or classify a patient into being awake or unconscious (binary classification models [6]) or as belonging to a variety of stages of consciousness (multi-label classification models [7–9]). Indeed, machine learning methods are well-suited to analyze time series of data such as electroencephalogram (EEG) signals, either in their raw form (with the use of RNNs and affiliated deep learning models) or following intensive preprocessing and feature extraction.

In many publications, researchers attempted to compare the performance of new algorithms to validated anesthesia depth monitors, such as the

bispectral index (BIS, Medtronic, USA) [5, 7, 10]. More recent papers have used AI to more directly estimate the depth of anesthesia from raw EEG, without feature extraction [6, 9, 11]. Besides EEG, a range of other signals have been investigated: heart rate variability [10], auditory evoked potentials [8], vital signs, and end-tidal carbon dioxide [12]. Another related application relates to the quantification of analgesia and degree of suppression of the analgesic response. This offers the possibility to differentiate analgesia from hypnosis and titrate opioids and sedative agents separately [13–16]. The performance (classification accuracy) of a few of these algorithms was tested against human performance, but there were only a handful of randomized trials comparing patient outcomes when patients were randomized into receiving either standard care or AI-augmented care [17–19]. What these trials showed was that automated adjustment of anesthetics drugs, when compared to human management, better maintained a given target within a narrow range and decreased overshooting and undershooting. The evidence of improved patient-centered outcomes such as neurocognitive recovery remains ambiguous [17–19].

Controlling Anesthesia Delivery with AI

A group of publications focus on the control of drug delivery for anesthesia, including sedative agents, opioids, muscle blockers, and related drugs.

These systems to automate anesthesia delivery require two inputs: (1) a measurement of the depth

of anesthesia (via the use of scores or surrogates of anesthetic depth based on EEG or clinical parameter measurements) and (2) a target level of sedation set by the user (Fig. 1).

Quite logically, several papers proposed to control the delivery of drugs in order to achieve a target value of BIS, either via simple pharmacokinetic models [20] or via machine learning, including reinforcement learning [21–23]. Earlier models proposed to use clinical signs to control the delivery of anesthetics [24, 25]. No paper described the use of direct EEG signal for anesthesia control [2].

Control systems that use machine learning have also been used to automate the delivery of neuromuscular blockade [26, 27], and these systems have also incorporated forecasting of drug pharmacokinetics to further improve the control of infusions of paralytics [28].

Interestingly, the level of maturity of some of these technologies was more advanced than other types of published machine learning research (see section below). The majority of publications described the development and the prospective validation of these tools, albeit in general in pilot studies on a very limited number of patients (around ten for most publications).

Perioperative Hemodynamic Optimization Assisted by AI

Perioperative hemodynamic optimization (a general approach for managing fluids, vasoressors, and inotropes, using advanced hemodynamic monitors, coupled with predefined treatment algorithms to achieve targeted physiological endpoints) has attracted considerable interest within the last three decades due to its potential to decrease postoperative morbidity and reduce hospital length of stay [29–33] and hospital costs [31, 34] and even to facilitate recovery [35] in patients undergoing major surgery when compared to routine care. The beneficial impact of hemodynamic optimization has been shown in multiple RCTs [33], meta-analyses [32], and quality improvement studies [36, 37],

all published in major scientific journals. This has led clinical societies in various countries to publish relevant national guidelines based on expert recommendations. These professional societies now consider this strategy as a standard of care and the best practice for perioperative hemodynamic optimization in high-risk surgical patients. However, despite the benefits, this concept is frequently not well applied at the bedside due to multiple factors including the fact that this practice is relatively time- and attention-consuming, which can jeopardize other clinical obligations and limit the “tight” control required of certain clinical scenarios [38, 39]. To overcome this issue, some teams around the world have developed automated systems to ease the titration of fluid and vasopressor in the perioperative setting. The application of closed-loop automation to better titrate fluid and vasopressor intraoperatively has the potential to benefit both the providers and patients. Figure 2 shows an operating room setup showing a prototype of a closed-loop vasopressor system during a cardiac surgery. The individual components of the system are indicated.

As monitoring systems continue to advance, inevitably the capabilities of autonomous systems will also advance. An ideal hemodynamic management system would be aware of not just vascular tone, but total body water and intravascular volume, the current infusion rates of anesthetic agents, the heart rate, pain levels, and cardiac stroke volume. Moreover, a holistic hemodynamic management system would have the means to modify all of these factors through pharmacologic or similar interventions. Given the complexity of these multiple-input multiple-output systems, the safety concerns, and the regulatory hurdles, it is likely that it will be some time before this type of system becomes commonplace. That said, independent automated systems have indeed been developed to autonomously manage anesthesia [40–42], fluid administration [43–45], and analgesics [46, 47], and recent studies have been published using more than one system simultaneously [18, 48, 49]. It is likely we will see much more such work in the years to come.

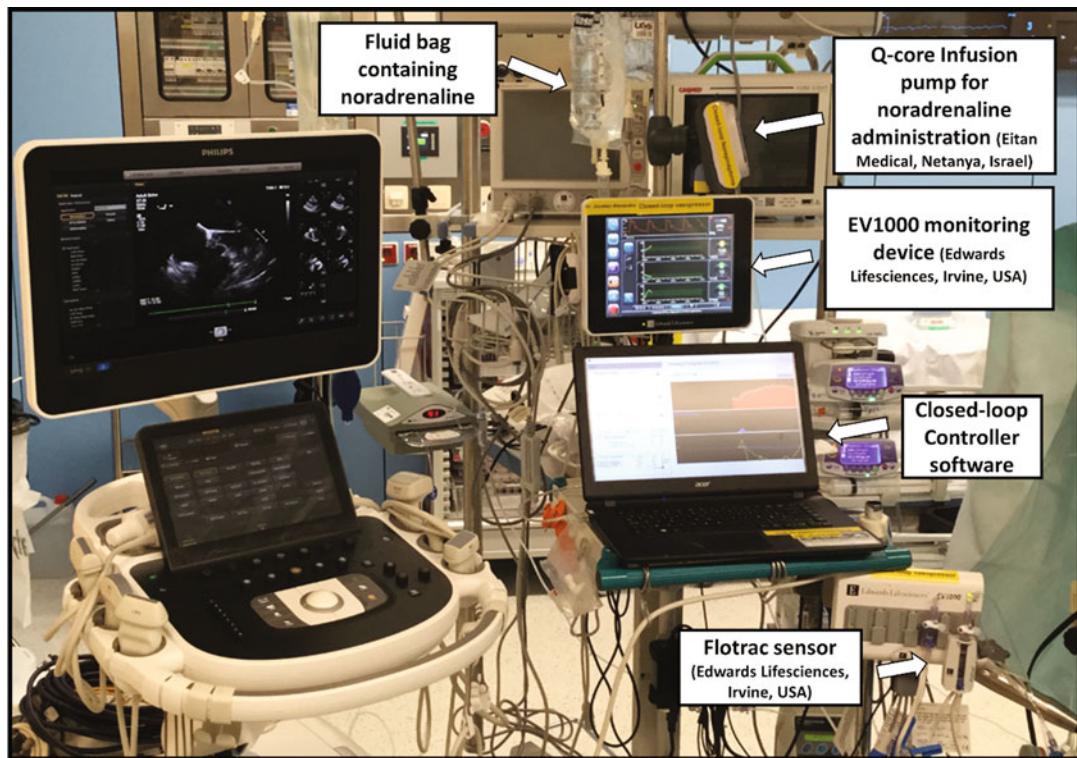


Fig. 2 Prototype closed-loop for vasopressor (CLV) system using off-the-shelf components in an operating room case at Erasme Hospital, Brussels, Belgium. Acer Laptop

running the algorithm software (connected to other components via RS232), FloTrac sensor providing clinical data, Q-core pump to deliver vasopressor

Automation for Fluid Therapy

Computer assistance of intravenous resuscitation was one of the first clinical interventions attempted. In the 1970s, a group in Texas developed a closed-loop system for fluid administration based on urine output [50, 51]. In the last 20 years, many measures have been used to direct fluid resuscitation, including mean arterial pressure [52, 53], spectroscopy [54], pulse pressure variation and stroke volume variation [55], or a combination thereof [56].

One of the challenges of fluid therapy is that unlike other measures like heart rate or blood pressure, “optimal” fluid status is neither clearly defined nor easily assessed [57, 58]. Despite these challenges, algorithms can nevertheless be designed to mimic with high fidelity the way that clinicians administer fluid, particularly when guided by established algorithms like goal-

directed fluid therapy [58]. While the question of “optimal” fluid strategy remains open, closed-loop systems may in the meantime ensure that protocols that have been shown effective may be implemented with higher consistency than when implemented manually [36, 59–61]. One author of this chapter review has published numerous human trials using this type of closed-loop goal-directed fluid therapy algorithm and has shown increased protocol adherence [43, 45, 62], decreased length of hospital stay [45], and decreased postoperative complications [63] in patients undergoing major surgery.

Several teams around the world are currently working on closed-loop fluid resuscitation algorithms [64–66]. Moreover, there have been early attempts to look at combining fluid resuscitation with other common resuscitation drugs like vaso-pressors [48, 67]. For the most part, many of these efforts are using independent, noncooperative

controllers, but showing this is safe is a necessary step in the progression of these technologies [48, 67]. Finally, closed-loop fluid delivery systems have also been used as an “unbiased interventionist” in studies examining the effects of different resuscitation fluids [44, 68], and therefore, such automated system can be a very promising tool to better design clinical studies and standardized treatment delivery. Today, two commercially available devices are widely available on the market (Concert and NeuroWave devices).

Automation for Vasopressor Titration

Several retrospective studies highlighted the key role of arterial pressure control in the operating room, and there is published evidence that the titration of vasopressors by hand is inefficient and inaccurate [69]. Thus, the development of a more effective titration system may be particularly clinically impactful and be the future for arterial pressure management in both surgical and intensive care unit patients [70, 71]. Vasopressor administration, having a single “controlled” physiologic variable readily measured (arterial pressure), provides a more direct target than fluid therapy; it is another intervention that earlier researchers attempted to automate [72–74]. However, there was little commercial development – likely due to a variety of factors, not the least of which is regulatory concerns and safety. More recently, some authors believe that such systems have a huge potential for the future [70, 71]. Available prototypes were developed for operating room and intensive care unit use [75–77], obstetrics and spinal-induced hypotension [78–80], and septic shock [81]. A recent paper published this year demonstrated that using such system resulted in a significantly lower incidence of intraoperative hypotension than routine care [82]. Moreover, Libert and colleagues in Paris developed and tested a closed-loop system that can coadminister vasopressors and fluids with good performance metrics [49]. Arterial pressure management, while complex and multifactorial, has great promise as a subject for automated management because of the known link to poor outcomes.

Automation for Inotrope Infusion

Cardiac function has not been studied much for pharmacological closed-loop systems. This may be a natural consequence of the properties of inotropic cardiac medications as well as difficulty in assessing ventricular contractility in an objective and continuous manner. Additionally, many cardiac drugs have nonlinear effects, and ceiling effects are common and often occur at low infusion rates. The obvious exception is the implantable cardiac pacemaker. The most modern pacemakers are quite advanced containing many layers of autonomous systems that combine several algorithms governing their function, but this is an electrical system and not a pharmacologic one [83].

Nevertheless, pharmacologic rate control of heart rate by automated systems has been done in cardiac stress tests [84]. Inotropes have also been studied in at least two published manuscripts [85, 86].

Automation for Vasodilator Infusion

Vasodilators, like vasopressors, have a directly measurable controlled variable and were thus also early autonomous system prototypes [87]. Because of the relative infrequent need for vasodilation drips in clinical care, however, and the risks of overtreatment, there has not been much progression of this research. Two studies did explore titration of vasodilators via closed-loop for neurosurgery [88] and cardiac surgery [89].

Machine Learning for Predicting Hypotension

Arterial pressure is one of the most important determinants of organ perfusion. Perioperative hypotension is frequent in patients undergoing surgery and in critically ill patients. The severity and duration of hypotension are both associated with tissue hypoperfusion and organ dysfunction [90]. Hypotension is mostly treated reactively after low arterial pressure values have already

occurred. However, prediction of hypotension before it becomes clinically apparent is now possible and allows the clinician to treat hypotension preemptively, thereby reducing the severity and duration of hypotension.

It has been clear for some years that machine learning can be used to predict hypotension 10–15 min before it clinically becomes apparent by analyzing subtle hemodynamic changes in the arterial waveform that precede clinical hypotension [90–92].

A hypotension prediction algorithm – resulting in a unit-less “hypotension prediction index” (HPI) – was proposed by Hatib and co-workers [93]. The machine learning algorithm was trained in a cohort of surgical and ICU patients and analyzes characteristics of the arterial pressure waveform. The HPI uses “a MAP of less than 65 mmHg for at least 1 minute” to define hypotension and can take values between 0 and 100, with higher numbers indicating a higher risk of hypotension. The initial validation study showed that the HPI was able to predict hypotension with a sensitivity of 88% and a specificity of 87% 15 min before a hypotensive event [93]. The predictive capabilities of the HPI algorithm were investigated in several studies. In a retrospective study including 23 patients undergoing cardiac and major vascular surgery, Ranucci et al. [94] reported that an HPI value of 85 had a sensitivity of 62% and a specificity of 78% (negative predictive value of 98%, positive predictive value of 13%) to predict hypotension 5–7 min before its clinical occurrence. In a retrospective study in 255 patients undergoing major surgery, Davies et al. [95] showed that the HPI predicted hypotension up to 15 min before its occurrence with a sensitivity of 81% and a specificity of 81%. Recently, the impact of HPI on the incidence, duration, and severity of hypotension was tested in clinical trials. In a single-center prospective interventional feasibility trial in patients undergoing primary hip arthroplasty, Schneck and colleagues [96] randomized 25 patients to goal-directed hemodynamic therapy based on HPI and 25 patients to routine management and additionally compared those groups to a historic control of 50 patients. Goal-directed hemodynamic

therapy based on HPI reduced the incidence, severity, and duration of intraoperative hypotension compared to the control groups. In a small single-center preliminary randomized trial including 68 adult patients scheduled for elective non-cardiac surgery under general anesthesia, Veelo et al. [97] showed that the use of the HPI in combination with a hemodynamic management protocol reduced the time-weighted average below a MAP of 65 mmHg (i.e., the area under a MAP of 65 mmHg divided by the duration of surgery) compared to standard care.

Very recently, Maheshwari and colleagues [98] demonstrated that HPI guidance did not reduce the duration and severity of hypotension in more than 200 patients undergoing noncardiac surgery. One of the hypotheses was that alerts were not followed by interventions, presumably due to short warning time and the complex treatment algorithm or because of clinicians ignoring the alert [99]. Indeed, correction of this new variable is still dependent upon clinician reactivity and intervention. Such indexes may reduce intraoperative hypotension but would be unlikely to reduce hypotensions the way an automated system can. Evidence of this was provided by Sessler and colleagues who showed, for example, that clinicians who had alerts on their hospital phone each time BIS, minimum alveolar concentration (MAC), or MAP was low (“triple low”) did not see any significant reduction in the rate or severity of these measures compared to those who did not [100]. The implication is that lack of awareness of these variables being out of range is not the principle limiting factor, but rather there is a limitation somewhere between awareness and correction. Lastly, we expect that in the future, the optimal hypotension threshold for a given patient might be determined using predictive analytics, further personalizing hemodynamic treatment.

Event and Risk Prediction with AI

A significant fraction of the literature in AI in anesthesiology focuses on predicting events or outcomes related to peri-anesthetic care. Schematically, these can be divided into:

1. *Preoperative tasks*, for example, for prediction of the class of anesthetic risk (using the American Society of Anesthesiologist classification) [101], detection of difficult intubation [102], or assistance in decision-making for the optimal method of anesthesia in pediatric surgery [103].
2. *Intra-operative optimization*, for the prediction of consciousness level in response to propofol boluses [104], prediction of length of surgery [105], or prediction of the rate of recovery from neuromuscular blockade [106].
3. *Postoperative care*, for the prediction of recovery or intrahospital length of stay [107] or for the prediction of postoperative complications such as acute kidney injury [108], delirium [109], deterioration and ICU admission [110], or mortality [111, 112].

While some of these models showed great performance, for example, in predicting postoperative mortality [111, 112], the accuracy of others was more modest. For example, Combes and colleagues used a hospital database containing extensive information on staffing, operating room use per procedure and staff, and postanesthesia care unit use with the electronic health record to train a neural network to predict the duration of an operation based on the team, type of operation, and a patient's relevant medical history; however, prediction accuracy of their models never exceeded 60% [105].

A recent paper published in *The New England Journal of Medicine* assessed for the first time on a large scale the benefit of an automated detection system for general clinical deterioration on the ward across 19 different hospitals, in a cohort of 548,383 non-ICU hospitalizations [110]. While nonrandomized, the authors were able to identify a significantly lower risk of 30-day mortality in hospitals where the AI system was in use (adjusted relative risk, 0.84; 95% confidence interval, 0.78–0.90; $P < 0.001$).

A few of these models were compared to human anesthesiologists. For example, neural networks were used to predict the hypnotic effect (as measured by BIS) of an induction bolus dose of propofol (sensitivity of 82.35%, specificity of 64.38%, and an area under the curve of 0.755) and

were found to exceed the average estimate of certified anesthesiologists (sensitivity, 20.64%; specificity, 92.51%; area under the curve of 0.5605) [104].

Other Applications of AI in Anesthesiology

AI techniques are also proposed in a range of various applications related to the field of anesthesiology, including in the control of mechanical ventilation, in the prediction of pain control requirements, or for ultrasound guidance.

In the field of mechanical ventilation, AI applications have been proposed for a range of tasks [113], including the control of the minute ventilation to achieve a target set end-tidal carbon dioxide [114] or automation of weaning from mechanical ventilation [115, 116].

With regard to pain management, machine learning has been used in various tasks such as identifying objective biomarkers for pain from fMRI [117] or skin conductance [13] or predicting postoperative pain levels or opioid requirements [118, 119]. Research is also ongoing in the domain of genetics of pain, with researchers trying to identify the genetic polymorphism associated with postoperative pain levels [120].

The use of convolutional neural networks for image analysis and classification is now widespread in the field of AI in radiology, with many clinical tools approved by the Food and Drug Administration for clinical use in the United States [121]. Several publications relate to the use of image processing or image segmentation and classification for needle guidance or identification of anatomical structures, including femoral vessels [122], vertebrae, and other anatomical landmarks for epidural placement [123, 124].

Technology Readiness Level of Published Applications

We assessed the degree of maturity of 137 AI applications listed in a 2020 review of the literature [2], after excluding research evaluating

Table 1 Degree of maturity of 137 AI applications in anesthesiology listed in a recent review of the literature [2]. OR, operating room; RCT, randomized controlled trial

Degree of maturity of technology	Depth of anesthesia	Control of anesthesia delivery	Event prediction	Ultrasound guidance	Pain mgmt.	OR logistics	Total
Retrospective model	32	3	41	6	6	1	89
Retrospective model with human comparison	4	0	4	1	0	0	9
Prospective testing on patients but no RCT	2	17	6	5	2	2	34
RCT	0	2	1	0	0	0	3
Evaluation of a certified device	1	0	1	0	0	0	2

existing certified medical devices and reviews (Table 1). What we found is that in agreement with other healthcare domains [125, 126], the maturity of the research is in general very low, with most publications (98/137 or 71%) relating to models that were developed and tested in a single dataset previously collected from a single institution. Only about a quarter of these publications (34/137) included some prospective clinical testing, in general on a very small number of patients (around 10). In this list of studies, we could only retrieve three examples of randomized controlled trials comparing standard of care to decisions supported by an AI tool.

Implications of AI for the Anesthesiologist

This review highlighted some current applications of AI in the field of anesthesiology. With progress in digitization of peri-anesthetic care, clinicians must consider how best to interpret the increasing amount of available data for the delivery of anesthetic and critical care. The application of AI technologies holds the potential to help clinicians maximize the clinical utility of the data that is captured electronically. AI has the potential to impact the practice of anesthesiology in many aspects, ranging from perioperative support to critical care delivery to outpatient pain management.

The predominant focus across most of these studies has been to investigate potential ways that

artificial intelligence can benefit the clinical practice of anesthesiology not through the replacement of the clinician but through augmentation of the anesthesiologist's workflow, decision-making, and some specific elements of clinical care. Thus, although AI is an expansive field, the results of this review demonstrated the literature's focus on the technology of machine learning, which is only one piece of the solution. The technology readiness level of the vast majority of current applications is very low, meaning that applications are nowhere near clinical deployment at scale. Only a handful of systems were tested in new settings or tested prospectively in randomized controlled trials against the standard of care. Of note, the question of the need and rationale for external validation and recalibration itself is debated [127].

Kuck and Johnson have formulated the three laws for automation in anesthesiology: (1) do no harm, (2) be transparent, and (3) reduce the cognitive workload [113]. Indeed, the dangers and risks of using these tools must not be forgotten. Sensor failure, unpredictable disturbances, and software errors remain issues in any automated system, and safety concepts need to be developed to ensure no harm is done to the patient [128]. Another issue to consider with predictive modeling is the discovery of patterns undetectable by the human mind that are not actually causing problems and never will (the problem of overdiagnosis) [129].

A further essential element is that very few systems were fitted into clinicians' workflow and

evaluated in relevant clinical environments. Besides their crude statistical performance, we anticipate that key aspects around usability, interface design, human factors, trust, and acceptability will heavily influence success or demise of these systems [125]. Finally, the scientific community has only scratched the surface on questions related to legal and ethical implications of using AI in anesthesiology and more generally in healthcare.

Because of these factors, we argue that to this day, the threat to human jobs remains distant. We do not anticipate that AI will master all aspects of clinical anesthesiology before a long time. While humans have intuition and the ability to extrapolate from their knowledge base to unknown situations, AI models can only draw conclusions from the data they were trained on [130]. However, as stated by Hashimoto and colleagues: “As more and more elements of clinical practice become digitized and accumulated into databases, and as fundamental algorithmic research keeps progressing, we may one day see the development of AI systems that have a more complete understanding of clinical phenomena and thus greater potential to deliver elements of anesthesia care autonomously” [2].

At this point in time, it is hard to predict what AI application will be the first to be used at scale or will have the most impact for patients and clinicians. From this analysis of the literature, we would cautiously bet on closed-loop systems designed to optimize hemodynamics or maintain general anesthesia and on automated deterioration alerting systems for ward patients.

Conclusion

AI in anesthesiology has the potential to support human decisions and improve outcomes, and some of the research has begun to bear fruit and was translated into commercially available systems. However, most of the current technology is in its early development phase, with very few examples of successful prospective deployment and testing, which should become the focus of teams and institutions involved in this research. In parallel and mirroring a phenomenon seen in

most medical specialties, gaining literacy in the field of AI technology will become inevitable for any practicing anesthesiologist.

References

- Barash PG, et al. Clinical anesthesia, 8e: print + Ebook with multimedia. 8th ed. LWW; 2017.
- Hashimoto DA, Witkowski E, Gao L, Meireles O, Rosman G. Artificial intelligence in anesthesiology current techniques, clinical applications, and limitations. *Anesthesiology*. 2020;132(2):379–94. <https://doi.org/10.1097/ALN.0000000000002960>.
- Veselis RA, Reinsel R, Sommer S, Carlon G. Use of neural network analysis to classify electroencephalographic patterns against depth of midazolam sedation in intensive care unit patients. *J Clin Monit*. 1991;7(3):259–67. <https://doi.org/10.1007/BF01619271>.
- Veselis RA, Reinsel R, Wronski M. Analytical methods to differentiate similar electroencephalographic spectra: neural network and discriminant analysis. *J Clin Monit*. 1993;9(4):257–67. <https://doi.org/10.1007/BF02886696>.
- Ortolani O, et al. EEG signal processing in anaesthesia. Use of a neural network technique for monitoring depth of anaesthesia. *Br J Anaesth*. 2002;88(5):644–8. <https://doi.org/10.1093/bja/88.5.644>.
- Mirsadeghi M, Behnam H, Shalbaf R, Jelveh Moghadam H. Characterizing awake and anesthetized states using a dimensionality reduction method. *J Med Syst*. 2016;40(1):13. <https://doi.org/10.1007/s10916-015-0382-4>.
- Benzy VK, Jasmin EA, Koshy RC, Amal F. Wavelet entropy based classification of depth of anesthesia. In: 2016 International conference on computational techniques in information and communication technologies (ICCTICT); 2016. p. 521–4. <https://doi.org/10.1109/ICCTICT.2016.7514635>.
- Nagaraj SB, et al. Patient-specific classification of ICU sedation levels from heart rate variability. *Crit Care Med*. 2017;45(7):e683–90. <https://doi.org/10.1097/CCM.0000000000002364>.
- Shalbaf A, Saffar M, Sleigh JW, Shalbaf R. Monitoring the depth of anesthesia using a new adaptive neurofuzzy system. *IEEE J Biomed Health Inform*. 2018;22(3):671–7. <https://doi.org/10.1109/JBHI.2017.2709841>.
- Liu Q, Ma L, Fan S-Z, Abbad MF, Shieh J-S. Sample entropy analysis for the estimating depth of anaesthesia through human EEG signal at different levels of unconsciousness during surgeries. *Peer J*. 2018;6:e4817. <https://doi.org/10.7717/peerj.4817>.
- Coşkun M, Guruler H, İstanbullu A, Peker M. Determining the appropriate amount of anesthetic gas using DWT and EMD combined with neural network. *J Med Syst*. 2015;39:1–10. <https://doi.org/10.1007/s10916-014-0173-3>.

12. Ranta SOV, Hynynen M, Räsänen J. Application of artificial neural networks as an indicator of awareness with recall during general anaesthesia. *J Clin Monit Comput.* 2002;17(1):53–60. <https://doi.org/10.1023/a:1015426015547>.
13. Ben-Israel N, Klinger M, Zuckerman G, Katz Y, Edry R. Monitoring the nociception level: a multi-parameter approach. *J Clin Monit Comput.* 2013;27(6):659–68. <https://doi.org/10.1007/s10877-013-9487-9>.
14. Daccache G, Jeanne M, Fletcher D. The analgesia nociception index: tailoring opioid administration. *Anesth Analg.* 2017;125(1):15–7. <https://doi.org/10.1213/ANE.0000000000002145>.
15. Janda M, et al. Design and implementation of a control system reflecting the level of analgesia during general anesthesia. *Biomed Tech (Berl).* 2013;58(1):1–11. <https://doi.org/10.1515/bmt-2012-0090>.
16. Upton HD, Ludbrook GL, Wing A, Sleigh JW. Intraoperative ‘analgesia nociception index’-guided fentanyl administration during sevoflurane anesthesia in lumbar discectomy and laminectomy: a randomized clinical trial. *Anesth Analg.* 2017;125(1):81–90. <https://doi.org/10.1213/ANE.0000000000001984>.
17. Cotoia A, et al. Effects of closed-loop intravenous anesthesia guided by bispectral index in adult patients on emergence delirium: a randomized controlled study. *Minerva Anestesiol.* 2018;84(4):437–46. <https://doi.org/10.23736/S0375-9393.17.11915-2>.
18. Joosten A, et al. Anesthetic management using multiple closed-loop systems and delayed neurocognitive recovery: a randomized controlled trial. *Anesthesiology.* 2020;132(2):253–66. <https://doi.org/10.1097/ALN.0000000000003014>.
19. Mahr N, et al. Postoperative neurocognitive disorders after closed-loop versus manual target controlled-infusion of propofol and remifentanil in patients undergoing elective major noncardiac surgery: the randomized controlled postoperative cognitive dysfunction-electroencephalographic-guided anesthetic administration trial. *Anesth Analg.* 2020; <https://doi.org/10.1213/ANE.0000000000005278>.
20. Absalom AR, Sutcliffe N, Kenny GN. Closed-loop control of anesthesia using bispectral index: performance assessment in patients undergoing major orthopedic surgery under combined general and regional anesthesia. *Anesthesiology.* 2002;96(1):67–73. <https://doi.org/10.1097/000000000000542-200201000-00017>.
21. Liu N, et al. Feasibility of closed-loop co-administration of propofol and remifentanil guided by the bispectral index in obese patients: a prospective cohort comparison. *BJA Br J Anesth.* 2015;114(4):605–14. <https://doi.org/10.1093/bja/aeu401>.
22. Lowery C, Faisal A. Towards efficient, personalized anesthesia using continuous reinforcement learning for propofol infusion control. 2013;1414–7. <https://doi.org/10.1109/NER.2013.6696208>.
23. Meskin N, Haddad W, Padmanabhan R. Closed-loop control of anesthesia and mean arterial pressure using reinforcement learning. *Biomed Signal Process Control.* 2015;22:54–64. <https://doi.org/10.1016/j.bspc.2015.05.013>.
24. Tsutsui T, Arita S. Fuzzy-logic control of blood pressure through enflurane anesthesia. *J Clin Monit.* 1994;10(2):110–7. <https://doi.org/10.1007/BF02886283>.
25. Zbinden AM, Feigenwinter P, Petersen-Felix S, Hacisalihzade S. Arterial pressure control with isoflurane using fuzzy logic. *Br J Anaesth.* 1995;74(1):66–72. <https://doi.org/10.1093/bja/74.1.66>.
26. Lendl M, Schwarz UH, Romeiser HJ, Unbehauen R, Georgieff M, Geldner GF. Nonlinear model-based predictive control of non-depolarizing muscle relaxants using neural networks. *J Clin Monit Comput.* 1999;15(5):271–8. <https://doi.org/10.1023/a:1009915105434>.
27. Shieh JS, Fan SZ, Chang LW, Liu CC. Hierarchical rule-based monitoring and fuzzy logic control for neuromuscular block. *J Clin Monit Comput.* 2000;16(8):583–92. <https://doi.org/10.1023/a:1012212516100>.
28. Motamed C, Devys J-M, Debaene B, Billard V. Influence of real-time Bayesian forecasting of pharmacokinetic parameters on the precision of a rocuronium target-controlled infusion. *Eur J Clin Pharmacol.* 2012;68(7):1025–31. <https://doi.org/10.1007/s00228-012-1236-3>.
29. Benes J, Giglio M, Brienza N, Michard F. The effects of goal-directed fluid therapy based on dynamic parameters on post-surgical outcome: a meta-analysis of randomized controlled trials. *Crit Care.* 2014;18(5):584. <https://doi.org/10.1186/s13054-014-0584-z>.
30. Corcoran T, Rhodes JEJ, Clarke S, Myles PS, Ho KM. Perioperative fluid management strategies in major surgery: a stratified meta-analysis. *Anesth Analg.* 2012;114(3):640–51. <https://doi.org/10.1213/ANE.0b013e318240d6eb>.
31. Ebm C, Cecconi M, Sutton L, Rhodes A. A cost-effectiveness analysis of postoperative goal-directed therapy for high-risk surgical patients. *Crit Care Med.* 2014;42(5):1194–203. <https://doi.org/10.1097/CCM.000000000000164>.
32. Hamilton MA, Cecconi M, Rhodes A. A systematic review and meta-analysis on the use of preemptive hemodynamic intervention to improve postoperative outcomes in moderate and high-risk surgical patients. *Anesth Analg.* 2011;112(6):1392–402. <https://doi.org/10.1213/ANE.0b013e3181eeaae5>.
33. Pearse RM, et al. Effect of a perioperative, cardiac output-guided hemodynamic therapy algorithm on outcomes following major gastrointestinal surgery: a randomized clinical trial and systematic review. *JAMA.* 2014;311(21):2181–90. <https://doi.org/10.1001/jama.2014.5305>.
34. Manecke GR, Asemota A, Michard F. Tackling the economic burden of postsurgical complications: would perioperative goal-directed fluid therapy help? *Crit Care.* 2014;18(5):566. <https://doi.org/10.1186/s13054-014-0566-1>.

35. Sun Y, Chai F, Pan C, Romeiser JL, Gan TJ. Effect of perioperative goal-directed hemodynamic therapy on postoperative recovery following major abdominal surgery-a systematic review and meta-analysis of randomized controlled trials. *Crit Care.* 2017;21(1):141. <https://doi.org/10.1186/s13054-017-1728-8>.
36. Cannesson M, et al. Perioperative goal-directed therapy and postoperative outcomes in patients undergoing high-risk abdominal surgery: a historical-prospective, comparative effectiveness study. *Crit Care.* 2015;19(1):261. <https://doi.org/10.1186/s13054-015-0945-2>.
37. Habicher M, et al. Implementation of goal-directed fluid therapy during hip revision arthroplasty: a matched cohort study. *Perioper Med.* 2016;5 <https://doi.org/10.1186/s13741-016-0056-x>.
38. Miller TE, Roche AM, Gan TJ. Poor adoption of hemodynamic optimization during major surgery: are we practicing substandard care? *Anesth Analg.* 2011;112(6):1274–6. <https://doi.org/10.1213/ANE.0b013e318218cc4f>.
39. Molliex S, et al. A multicentre observational study on management of general anaesthesia in elderly patients at high-risk of postoperative adverse outcomes. *Anaesth Crit Care Pain Med.* 2019;38(1):15–23. <https://doi.org/10.1016/j.accpm.2018.05.012>.
40. Kong E, Nicolaou N, Vizcaychipi MP. Hemodynamic stability of closed-loop anesthesia systems: a systematic review. *Minerva Anestesiol.* 2020;86(1):76–87. <https://doi.org/10.23736/S0375-9393.19.13927-2>.
41. Guen MLE, Liu N, Chazot T, Fischler M. Closed-loop anesthesia. *Minerva Anestesiol.* 2016;82(5):573–81.
42. West N, et al. Design and evaluation of a closed-loop anesthesia system with robust control and safety system. *Anesth Analg.* 2018;127(4):883–94. <https://doi.org/10.1213/ANE.0000000000002663>.
43. Joosten A, Huynh T, Suehiro K, Canales C, Cannesson M, Rinehart J. Goal-directed fluid therapy with closed-loop assistance during moderate risk surgery using noninvasive cardiac output monitoring: a pilot study. *Br J Anaesth.* 2015;114(6):886–92. <https://doi.org/10.1093/bja/aev002>.
44. Joosten A, et al. Crystalloid versus colloid for intraoperative goal-directed fluid therapy using a closed-loop system: a randomized, double-blinded, controlled trial in major abdominal surgery. *Anesthesiology.* 2018;128(1):55–66. <https://doi.org/10.1097/ALN.0000000000001936>.
45. Rinehart J, et al. Closed-loop assisted versus manual goal-directed fluid therapy during high-risk abdominal surgery: a case-control study with propensity matching. *Crit Care.* 2015;19(1) <https://doi.org/10.1186/s13054-015-0827-7>.
46. Liu N, et al. Feasibility of closed-loop titration of propofol and remifentanil guided by the spectral M-entropy monitor. *Anesthesiology.* 2012;116(2): 286–95. <https://doi.org/10.1097/ALN.0b013e318242ad4f>.
47. Luginbühl M, Bieniok C, Leibundgut D, Wymann R, Gentilini A, Schnider TW. Closed-loop control of mean arterial blood pressure during surgery with alfentanil: clinical evaluation of a novel model-based predictive controller. *Anesthesiology.* 2006;105(3):462–70. <https://doi.org/10.1097/00000542-200609000-00008>.
48. Joosten A, et al. Fully automated anesthesia and fluid management using multiple physiologic closed-loop systems in a patient undergoing high-risk surgery. *Case Rep.* 2016;7(12):260–5. <https://doi.org/10.1213/XAA.000000000000405>.
49. Libert N, et al. Performance of closed-loop resuscitation of haemorrhagic shock with fluid alone or in combination with norepinephrine: an experimental study. *Ann Intensive Care.* 2018;8 <https://doi.org/10.1186/s13613-018-0436-0>.
50. Bowman RJ, Westenskow DR. A microcomputer-based fluid infusion system for the resuscitation of burn patients. *IEEE Trans Biomed Eng.* 1981;28(6): 475–9. <https://doi.org/10.1109/TBME.1981.324822>.
51. DeBey RK, Westenskow DR, Jordan WS, McJames SW. A urine based control system for fluid infusion. *Biomed Sci Instrum.* 1987;23:195–8.
52. Blankenship HB, Wallace FD, Pacifico AD. Clinical application of closed-loop postoperative autotransfusion. *Med Prog Technol.* 1990;16(1–2):89–93.
53. Hoskins SL, et al. Closed-loop resuscitation of burn shock. *J Burn Care Res.* 2006;27(3):377–85. <https://doi.org/10.1097/01.BCR.0000216512.30415.78>.
54. Chaisson NF, Kirschner RA, Deyo DJ, Lopez JA, Prough DS, Kramer GC. Near-infrared spectroscopy-guided closed-loop resuscitation of hemorrhage. *J Trauma.* 2003;54(5 Suppl):S183–92. <https://doi.org/10.1097/01.TA.0000064508.11512.28>.
55. Rinehart J, et al. Evaluation of a novel closed-loop fluid-administration system based on dynamic predictors of fluid responsiveness: an in silico simulation study. *Crit Care.* 2011;15(6):R278. <https://doi.org/10.1186/cc10562>.
56. Rinehart J, Chung E, Canales C, Cannesson M. Intraoperative stroke volume optimization using stroke volume, arterial pressure, and heart rate: closed-loop (learning intravenous resuscitator) versus anesthesiologists. *J Cardiothorac Vasc Anesth.* 2012;26(5): 933–9. <https://doi.org/10.1053/j.jvca.2012.05.015>.
57. Cannesson M. Arterial pressure variation and goal-directed fluid therapy. *J Cardiothorac Vasc Anesth.* 2010;24(3):487–97. <https://doi.org/10.1053/j.jvca.2009.10.008>.
58. Rinehart J, Liu N, Alexander B, Cannesson M. Review article: closed-loop systems in anesthesia: is there a potential for closed-loop fluid management and hemodynamic optimization? *Anesth Analg.* 2012;114(1):130–43. <https://doi.org/10.1213/ANE.0b013e318230e9e0>.
59. Ramsingh DS, Sanghvi C, Gamboa J, Cannesson M, Applegate RL. Outcome impact of goal directed fluid therapy during high risk abdominal surgery in low to moderate risk patients: a randomized controlled trial. *J Clin Monit Comput.* 2013;27(3):249–57. <https://doi.org/10.1007/s10877-012-9422-5>.
60. Ripollés-Melchor J, et al. Perioperative goal-directed hemodynamic therapy in noncardiac surgery: a

- systematic review and meta-analysis. *J Clin Anesth.* 2016;28:105–15. <https://doi.org/10.1016/j.jclinane.2015.08.004>.
61. Rollins KE, Lobo DN. Intraoperative goal-directed fluid therapy in elective major abdominal surgery: a meta-analysis of randomized controlled trials. *Ann Surg.* 2016;263(3):465–76. <https://doi.org/10.1097/SLA.0000000000001366>.
62. Joosten A, et al. Implementation of closed-loop-assisted intra-operative goal-directed fluid therapy during major abdominal surgery: a case-control study with propensity matching. *Eur J Anaesthesiol.* 2018;35(9):650–8. <https://doi.org/10.1097/EJA.0000000000000827>.
63. Joosten A, et al. Practical impact of a decision support for goal-directed fluid therapy on protocol adherence: a clinical implementation study in patients undergoing major abdominal surgery. *J Clin Monit Comput.* 2019;33(1):15–24. <https://doi.org/10.1007/s10877-018-0156-x>.
64. Gholami B, Haddad WM, Bailey JM, Geist B, Ueyama Y, Muir WW. A pilot study evaluating adaptive closed-loop fluid resuscitation during states of absolute and relative hypovolemia in dogs. *J Vet Emerg Crit Care San Antonio.* 2018; 436–46.
65. Jin X, Bighamian R, Hahn J-O. Development and in silico evaluation of a model-based closed-loop fluid resuscitation control algorithm. *IEEE Trans Biomed Eng.* 2018; <https://doi.org/10.1109/TBME.2018.2880927>.
66. Lilot M, et al. Comparison of cardiac output optimization with an automated closed-loop goal-directed fluid therapy versus non standardized manual fluid administration during elective abdominal surgery: first prospective randomized controlled trial. *J Clin Monit Comput.* 2018;32(6):993–1003. <https://doi.org/10.1007/s10877-018-0106-7>.
67. Joosten A, et al. Feasibility of fully automated hypnosis, analgesia, and fluid management using 2 independent closed-loop systems during major vascular surgery: a pilot study. *Anesth Analg.* 2019;128(6): e88–92. <https://doi.org/10.1213/ANE.0000000000003433>.
68. Joosten A, et al. Long-term impact of crystalloid versus colloid solutions on renal function and disability-free survival after major abdominal surgery. *Anesthesiology.* 2019;130(2):227–36. <https://doi.org/10.1097/ALN.0000000000002501>.
69. Rinehart J, et al. Blood pressure variability in surgical and intensive care patients: is there a potential for closed-loop vasopressor administration? *Anaesth Crit Care Pain Med.* 2019;38(1):69–71. <https://doi.org/10.1016/j.accpm.2018.11.009>.
70. Joosten A, Rinehart J. Part of the steamroller and not part of the road: better blood pressure management through automation. *Anesth Analg.* 2017;125(1): 20–2. <https://doi.org/10.1213/ANE.0000000000002201>.
71. Michard F, Liu N, Kurz A. The future of intraoperative blood pressure management. *J Clin Monit Comput.* 2018;32(1):1–4. <https://doi.org/10.1007/s10877-017-9989-y>.
72. Mason DG, Packer JS, Cade JF, McDonald RD. Closed-loop management of blood pressure in critically ill patients. *Australas Phys Eng Sci Med.* 1985;8(4):164–7.
73. Packer JS, Mason DG, Cade JF, McKinley SM. An adaptive controller for closed-loop management of blood pressure in seriously ill patients. *IEEE Trans Biomed Eng.* 1987;34(8):612–6. <https://doi.org/10.1109/tbme.1987.326072>.
74. Potter DR, Moyle JT, Lester RJ, Ware RJ. Closed loop control of vasoactive drug infusion. A preliminary report. *Anaesthesia.* 1984;39(7):670–7. <https://doi.org/10.1111/j.1365-2044.1984.tb06476.x>.
75. Joosten A, et al. Automated titration of vasopressor infusion using a closed-loop controller: in vivo feasibility study using a swine model. *Anesthesiology.* 2019;130(3):394–403. <https://doi.org/10.1097/ALN.0000000000002581>.
76. Rinehart J, Ma M, Calderon M-D, Cannesson M. Feasibility of automated titration of vasopressor infusions using a novel closed-loop controller. *J Clin Monit Comput.* 2018;32(1):5–11. <https://doi.org/10.1007/s10877-017-9981-6>.
77. Rinehart J, Joosten A, Ma M, Calderon M-D, Cannesson M. Closed-loop vasopressor control: in-silico study of robustness against pharmacodynamic variability. *J Clin Monit Comput.* 2019;33(5): 795–802. <https://doi.org/10.1007/s10877-018-0234-0>.
78. Ngan Kee WD, Tam YH, Khaw KS, Ng FF, Critchley LA, Karmakar MK. Closed-loop feedback computer-controlled infusion of phenylephrine for maintaining blood pressure during spinal anaesthesia for caesarean section: a preliminary descriptive study. *Anaesthesia.* 2007;62(12):1251–6. <https://doi.org/10.1111/j.1365-2044.2007.05257.x>.
79. Ngan Kee WD, Khaw KS, Tam Y-H, Ng FF, Lee SW. Performance of a closed-loop feedback computer-controlled infusion system for maintaining blood pressure during spinal anaesthesia for caesarean section: a randomized controlled comparison of norepinephrine versus phenylephrine. *J Clin Monit Comput.* 2017;31(3):617–23. <https://doi.org/10.1007/s10877-016-9883-z>.
80. Ngan Kee WD, Tam Y-H, Khaw KS, Ng FF, Lee SWY. Closed-loop feedback computer-controlled phenylephrine for maintenance of blood pressure during spinal anaesthesia for cesarean delivery: a randomized trial comparing automated boluses versus infusion. *Anesth Analg.* 2017;125(1):117–23. <https://doi.org/10.1213/ANE.0000000000001974>.
81. Merouani M, et al. Norepinephrine weaning in septic shock patients by closed loop control based on fuzzy logic. *Crit Care.* 2008;12(6):R155. <https://doi.org/10.1186/cc7149>.
82. Joosten A, et al. Automated closed-loop versus manually controlled norepinephrine infusion in patients undergoing intermediate- to high-risk abdominal surgery: a randomised controlled trial. *Br J Anaesth.* 2020; <https://doi.org/10.1016/j.bja.2020.08.051>.

83. Tjong FVY, Reddy VY. Permanent leadless cardiac pacemaker therapy: a comprehensive review. *Circulation*. 2015;135(15):1458–70. <https://doi.org/10.1161/CIRCULATIONAHA.116.025037>.
84. Sarabadiani Tafreshi A, Klamroth-Marganska V, Nussbaumer S, Riener R. Real-time closed-loop control of human heart rate and blood pressure. *IEEE Trans Biomed Eng*. 2015;62(5):1434–42. <https://doi.org/10.1109/TBME.2015.2391234>.
85. Osswald S, et al. Closed-loop stimulation using intracardiac impedance as a sensor principle: correlation of right ventricular dP/dtmax and intracardiac impedance during dobutamine stress test. *Pacing Clin Electrophysiol PACE*. 2000;23(10 Pt 1):1502–8. <https://doi.org/10.1046/j.1460-9592.2000.01502.x>.
86. Uemura K, Kawada T, Zheng C, Sugimachi M. Less invasive and inotrope-reduction approach to automated closed-loop control of hemodynamics in decompensated heart failure. *IEEE Trans Biomed Eng*. 2016;63(8):1699–708. <https://doi.org/10.1109/TBME.2015.2499782>.
87. Hammond JJ, Kirkendall WM, Calfee RV. Hypertensive crisis managed by computer controlled infusion of sodium nitroprusside: a model for the closed loop administration of short acting vasoactive agents. *Comput Biomed Res Int J*. 1979;12(2):97–108. [https://doi.org/10.1016/0010-4809\(79\)90008-9](https://doi.org/10.1016/0010-4809(79)90008-9).
88. Mackenzie AF, Colvin JR, Kenny GN, Bisset WI. Closed loop control of arterial hypertension following intracranial surgery using sodium nitroprusside. A comparison of intra-operative halothane or isoflurane. *Anesthesia*. 1993;48(3):202–4. <https://doi.org/10.1111/j.1365-2044.1993.tb06901.x>.
89. Bednarski P, Siclari F, Voigt A, Demertzis S, Lau G. Use of a computerized closed-loop sodium nitroprusside titration system for antihypertensive treatment after open heart surgery. *Crit Care Med*. 1990;18(10):1061–5. <https://doi.org/10.1097/00003246-199010000-00002>.
90. Pinsky MR. Complexity modeling: identify instability early. *Crit Care Med*. 2010;38(10 Suppl):S649–55. <https://doi.org/10.1097/CCM.0b013e3181f24484>.
91. Hravnak M, et al. Artifact patterns in continuous noninvasive monitoring of patients. *Intensive Care Med*. 2013;39(Suppl 2):S405.
92. Pinsky MR, Dubrawski A. Gleaning knowledge from data in the intensive care unit. *Am J Respir Crit Care Med*. 2014;190(6):606–10. <https://doi.org/10.1164/rccm.201404-0716CP>.
93. Hatib F, et al. Machine-learning algorithm to predict hypotension based on high-fidelity arterial pressure waveform analysis. *Anesthesiology*. 2018;129(4):663–74. <https://doi.org/10.1097/ALN.0000000000002300>.
94. Ranucci M, Barile L, Ambrogi F, Pistuddi V, Surgical and Clinical Outcome Research (SCORE) Group. Dissemination and calibration properties of the hypotension probability indicator during cardiac and vascular surgery. *Minerva Anestesiol*. 2019;85(7):724–30. <https://doi.org/10.23736/S0375-9393.18.12620-4>.
95. Davies SJ, Vistisen ST, Jian Z, Hatib F, Scheeren TWL. Ability of an arterial waveform analysis-derived hypotension prediction index to predict future hypotensive events in surgical patients. *Anesth Analg*. 2020;130(2):352–9. <https://doi.org/10.1213/ANE.0000000000004121>.
96. Schneck E, et al. Hypotension prediction index based protocolized haemodynamic management reduces the incidence and duration of intraoperative hypotension in primary total hip arthroplasty: a single centre feasibility randomised blinded prospective interventional trial. *J Clin Monit Comput*. 2020;34(6):1149–58. <https://doi.org/10.1007/s10877-019-00433-6>.
97. Wijnberge M, et al. Effect of a machine learning-derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the HYPE randomized clinical trial. *JAMA*. 2023;323(11):1052–60. <https://doi.org/10.1001/jama.2020.0592>.
98. Maheshwari K, et al. Hypotension prediction index for prevention of hypotension during moderate- to high-risk noncardiac surgery: a pilot randomized trial. *Anesthesiology*. 2020;133(6):1214–22. <https://doi.org/10.1097/ALN.0000000000003557>.
99. Joosten A, Rinehart J, Cannesson M. Perioperative goal directed therapy: evidence and compliance are two sides of the same coin. *Rev Esp Anestesiol Reanim*. 2015;62(4):181–3. <https://doi.org/10.1016/j.redar.2015.01.012>.
100. Sessler DI, et al. Triple-low alerts do not reduce mortality: a real-time randomized trial. *Anesthesiology*. 2019;130(1):72–82. <https://doi.org/10.1097/ALN.0000000000002480>.
101. Zhang L, Fabbri D, Lasko TA, Ehrenfeld JM, Wanderer JP. A system for automated determination of perioperative patient acuity. *J Med Syst*. 2018;42(7):123. <https://doi.org/10.1007/s10916-018-0977-7>.
102. Moustafa MA, El-Metainy S, Mahar K, Abdellmagied EM. Defining difficult laryngoscopy findings by using multiple parameters: a machine learning approach. *Egypt J Anaesth*. 2017;33(2):153–8. <https://doi.org/10.1016/j.ejga.2017.02.002>.
103. Hancerliogullari G, Hancerliogullari KO, Koksalmis E. The use of multi-criteria decision making models in evaluating anaesthesia method options in circumcision surgery. *BMC Med Inform Decis Mak*. 2017;17(1):14. <https://doi.org/10.1186/s12911-017-0409-5>.
104. Lin C-S, Li Y-C, Mok MS, Wu C-C, Chiu H-W, Lin Y-H. Neural network modeling to predict the hypnotic effect of propofol bolus induction. *Proc AMIA Symp*. 2002:450–3.
105. Combes C, Meskens N, Rivat C, Vandamme J-P. Using a KDD process to forecast the duration of surgery. *Int J Prod Econ*. 2008;112(1):279–93. <https://doi.org/10.1016/j.ijpe.2006.12.068>.

106. Santanen OAP, Svartling N, Haasio J, Paloheimo MPJ. Neural nets and prediction of the recovery rate from neuromuscular block. *Eur J Anaesthesiol.* 2003;20(2):87–92. <https://doi.org/10.1017/s0265021503000164>.
107. Kumar A, Anjomshoa H. A two-stage model to predict surgical patients' lengths of stay from an electronic patient database. *IEEE J Biomed Health Inform.* 2019;23(2):848–56. <https://doi.org/10.1109/JBHI.2018.2819646>.
108. Lei VJ, et al. Risk stratification for postoperative acute kidney injury in major noncardiac surgery using preoperative and intraoperative data. *JAMA Netw Open.* 2019;2(12) <https://doi.org/10.1001/jamanetworkopen.2019.16921>.
109. Galyfos GC, Geropapas GE, Sianou A, Sigala F, Filis K. Risk factors for postoperative delirium in patients undergoing vascular surgery. *J Vasc Surg.* 2017;66(3):937–46. <https://doi.org/10.1016/j.jvs.2017.03.439>.
110. Escobar GJ, Liu VX, Schuler A, Lawson B, Greene JD, Kipnis P. Automated identification of adults at risk for in-hospital clinical deterioration. *N Engl J Med.* 2020;383(20):1951–60. <https://doi.org/10.1056/NEJMsa2001090>.
111. Lee CK, Hofer I, Gabel E, Baldi P, Cannesson M. Development and validation of a deep neural network model for prediction of postoperative in-hospital mortality. *Anesthesiology.* 2018;129(4):649–62. <https://doi.org/10.1097/ALN.0000000000002186>.
112. Wong DJN, et al. Developing and validating subjective and objective risk-assessment measures for predicting mortality after major surgery: an international prospective cohort study. *PLoS Med.* 2020;17(10):e1003253. <https://doi.org/10.1371/journal.pmed.1003253>.
113. von Platen P, Pomprapa A, Lachmann B, Leonhardt S. The dawn of physiological closed-loop ventilation – a review. *Crit Care.* 2020;24(1):121. <https://doi.org/10.1186/s13054-020-2810-1>.
114. Schäublin J, Derighetti M, Feigenwinter P, Petersen-Felix S, Zbinden AM. Fuzzy logic control of mechanical ventilation during anaesthesia. *Br J Anaesth.* 1996;77(5):636–41. <https://doi.org/10.1093/bja/77.5.636>.
115. N. Prasad, L.-F. Cheng, C. Chivers, M. Draugelis, and B. E. Engelhardt (2017) A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. ArXiv170406300 Cs [Online]. <http://arxiv.org/abs/1704.06300>.
116. Schädler D, et al. A knowledge- and model-based system for automated weaning from mechanical ventilation: technical description and first clinical application. *J Clin Monit Comput.* 2014;28(5):487–98. <https://doi.org/10.1007/s10877-013-9489-7>.
117. Mackey S, Greely HT, Martucci KT. Neuroimaging-based pain biomarkers: definitions, clinical and research applications, and evaluation frameworks to achieve personalized pain medicine. *Pain Rep.* 2019;4(4) <https://doi.org/10.1097/PR9.0000000000000762>.
118. Gram M, et al. Prediction of postoperative opioid analgesia using clinical-experimental parameters and electroencephalography. *Eur J Pain.* 2017;21(2):264–77. <https://doi.org/10.1002/ejp.921>.
119. Hu Y-J, Ku T-H, Jan R-H, Wang K, Tseng Y-C, Yang S-F. Decision tree-based learning to predict patient controlled analgesia consumption and readjustment. *BMC Med Inform Decis Mak.* 2012;12:131. <https://doi.org/10.1186/1472-6947-12-131>.
120. Olesen AE, Grønlund D, Gram M, Skorpen F, Drewes AM, Klepstad P. Prediction of opioid dose in cancer pain patients using genetic profiling: not yet an option with support vector machine learning. *BMC Res Notes.* 2018;11(1):78. <https://doi.org/10.1186/s13104-018-3194-z>.
121. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25(1):44. <https://doi.org/10.1038/s41591-018-0300-7>.
122. Smistad E, Løvstakken L. Vessel detection in ultrasound images using deep convolutional neural networks. In: Deep learning and data labeling for medical applications. Cham: Springer; 2016. p. 30–8. https://doi.org/10.1007/978-3-319-46976-8_4.
123. Hetherington J, Lessoway V, Gunka V, Abolmaesumi P, Rohling R. SLIDE: automatic spine level identification system using a deep convolutional neural network. *Int J Comput Assist Radiol Surg.* 2017;12(7):1189–98. <https://doi.org/10.1007/s11548-017-1575-8>.
124. Pesteie M, Lessoway V, Abolmaesumi P, Rohling RN. Automatic localization of the needle target for ultrasound-guided epidural injections. *IEEE Trans Med Imaging.* 2018;37(1):81–92. <https://doi.org/10.1109/TMI.2017.2739110>.
125. Komorowski M. Artificial intelligence in intensive care: are we there yet? *Intensive Care Med.* 2019;45(9):1298–300. <https://doi.org/10.1007/s00134-019-05662-6>.
126. Komorowski M. Clinical management of sepsis can be improved by artificial intelligence: yes. *Intensive Care Med.* 2020;46(2):375–7. <https://doi.org/10.1007/s00134-019-05898-2>.
127. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health.* 2020;2(9):e489–92. [https://doi.org/10.1016/S2589-7500\(20\)30186-2](https://doi.org/10.1016/S2589-7500(20)30186-2).
128. McDermid JA, Jia Y, Habli I. Towards a framework for safety assurance of autonomous systems. *Artif Intell Safety.* 2019;11:2019. <http://eprints.whiterose.ac.uk/150187/>. Accessed 18 Nov 2020
129. Komorowski M, Celi LA. Will artificial intelligence contribute to overuse in healthcare? *Crit Care Med.* 2017;45(5):912–3. <https://doi.org/10.1097/CCM.0000000000002351>.
130. Gambus P, Shafer SL. Artificial intelligence for everyone. *Anesthesiology.* 2018;128(3):431–3. <https://doi.org/10.1097/ALN.0000000000001984>.



Artificial Intelligence in Critical Care 106

The Path from Promise to Practice

Alfredo Vellido and Vicent Ribas

Contents

Introduction	1470
Interpretation, Explanation, Pipelines, and Guidelines	1471
Where at the ICU Should We Apply AI and ML?	1472
What Sort of AI and ML Can We Apply at the ICU?	1474
A Further Few Things About the Use of AI and ML in Medicine that Merit Discussion	1475
Conclusions	1476
References	1476

Abstract

The hectic domain of critical care is, arguably, one of the most demanding instances of multi-disciplinary medical decision-making. It is also a domain infused with monitoring and life-

sustaining technologies that leave behind the type of digital data trail that is ideal for the deployment of artificial intelligence and, particularly, machine learning methods and analytical pipelines. These data analysis methods should provide medical decision support to intensivists, but there is yet no clear framework or standard set of guidelines on how to do so. In this chapter, we aim to provide readers with information on the current landscape of applications of machine learning in critical care and also with a clear outline of some of the many challenges yet to overcome to guarantee the success of such applications, including the problem of model interpretability and explainability, the technology compliance with current legislation, or the education of medical practitioners in the area of data analytics using open-source software.

Supplementary Information: The online version of this chapter (https://doi.org/10.1007/978-3-030-64573-1_174) contains supplementary material, which is available to authorized users.

A. Vellido (✉)
Universitat Politècnica de Catalunya, Barcelona, Spain
Intelligent Data Science and Artificial Intelligence (IDEAI-UPC) Research Center, Barcelona, Spain
e-mail: a.vellido@cs.upc.edu

V. Ribas
Eurecat, Barcelona, Spain
e-mail: vicent.ribas@eurecat.org

Keywords

Artificial intelligence · Machine learning · Deep learning · Critical care · Intensive care unit

Introduction

The critical care domain and, more specifically, the current intensive care unit (ICU) in its different forms (neonatal, pediatric, geriatric, coronary, etc.) rely heavily on technology. All sorts of physiological data from patients at the ICU are gathered in order to inform physicians in a hectic environment in which timely delivered information is paramount to save people's lives. After all, patients in this environment are technologically dependent on life-sustaining devices such as infusion pumps, mechanical ventilators, or catheters, to name a few.

The patient's data captured by the bedside are conditioned by changes in patient status and may consist of electronic health record and monitoring devices data, including, among others, fluid intake and patient output, laboratory blood draw analyses, medical images, demographics, etc. Most of these measurements are electronically stored, generating a digital signature over time of the patients during their stay at the ICU.

The real-time information gathered in this way can be used by the physician as a guide to inform clinical decisions concerning diagnosis, prognosis, and course of treatment. Nevertheless, this information cannot be considered to be knowledge as such, just yet. Knowledge only resides in the physician's biological intelligence, working against a background of expert experience, provided by training and practice. Needless to say, that knowledge is bounded by the limitations of human experience: a physician can only operationalize the knowledge obtained from personal practice, which is augmented by the evidence-based medical protocols and guidelines that result from the accumulated experience of the profession for a given problem.

All the physiological measurements and further information gathered at the point of care conform a true digital signature of the ICU patient. Any clinical environment can thus gather thousands upon thousands of those signatures over time, building databases whose accumulated information is well beyond the reach of the "information-processing capabilities" of any individual medical expert. Extracting useful and actionable knowledge out of these data that go beyond the experience of an individual physician to reach a whole patient population is, instead, the domain of pattern recognition approaches. This is why machine learning (ML) and artificial intelligence (AI) at large are called to play an important role in the ICU: as techniques that can extract knowledge from large, heterogeneous, and complex databases to feed the human knowledge base for specific problems in critical care. Note though that the realization of large inter-center, international critical care databases faces severe challenges. Due to methodological differences in the way critical care is delivered in different medical centers, cultural differences in critical care documentation, and normative disparity, "the development of a system for crosstalk between critical care databases would be a major engineering feat," as argued in [1].

This chapter aims to provide the reader with a landscape of the many opportunities offered by ML and AI to the critical care community, both for research and for point-of-care applications. This will include commentary from a twofold perspective: that of the ML/AI techniques and families of techniques applied in this domain and that of the type of critical care problems that are currently being investigated using ML/AI-based approaches. Beyond that, this chapter addresses several issues concerning the application of ML/AI in critical care whose importance is only now becoming clear in medical applications in general. These issues include the need of interpretable and explainable models; the relevance of designing complete analytical pipelines that are compatible with medical guidelines applied at the point-of-care; and the required compliance of ML/AI methods with boundaries set by ethics and law.

Interpretation, Explanation, Pipelines, and Guidelines

Data dependence is bound to increase in healthcare and medical practice in general and, in particular, in the ICU. As stated in the introduction, this could be seen as an environment ready for data science in general, and ML/AI in particular, to thrive. The fact, though, is that even in this apparently perfect match between an information-rich medical environment and the broad availability of ML/AI data analysis methods, these approaches are not yet universally accepted by the medical community and far from being widely adopted, despite of their maturity, after decades of research and development.

It is argued in this section that the main reasons behind this apparent contradiction have little to do with technological maturity or readiness. Instead, they mostly have to do with difficulties of adoption and implementation challenges that, unless resolved, will not allow ML/AI methods to be adopted in routine medical practice beyond a reduced number of niche applications. One of these challenges is the inherent lack of interpretability and explainability of many ML/AI techniques [2], or, as eloquently expressed in [3], “the need to open the machine learning black box.” This “black box syndrome” has been discussed for decades in the medical domain [4]. Unfortunately, this domain challenge has not entered the mainstream discussion until recently [5], at least partially because of the huge success enjoyed by the current iteration of artificial neural network (ANN) techniques defined as deep learning (DL), which have found their way into biomedicine and healthcare [2, 6] and which are often referred to as an extreme case of model opacity.

This problem should clearly resonate with critical care practitioners: an ML/AI-based system may yield timely results that, despite their accuracy, might not be amenable of comprehensible explication. In an environment prone to swift decision-making such as the ICU, this might pose an insurmountable barrier for adoption, as the medical expert could not trust to implement a decision that cannot be either comprehended or

explained to either the patient or to other medical experts. Despite this, it must be noted that plenty of research has been devoted to imbue ML/AI methods with interpretability qualities, so that complex predictive models can be explained. Most approaches aim to somehow replicate human interpretation procedures, for instance, mimicking the performance of complex models using simpler and therefore more interpretable one; using visual analytics to provide intuitive insights [7]; rule extraction to make model interpretation akin to medical guidelines; or semantic representations that keep interpretation closer to human natural language.

This attempt to provide complex models with interpretability that is compatible with human reasoning brings another side to the argument, which has been broached in [8]: the need to integrate the available human expert knowledge into the ML/AI models. This means that formal frameworks for machine-human interaction that put the medical expert at the core of the process to achieve interpretability and explainability are compulsory [9] and that they should be able to guarantee a fluid exchange of knowledge between the ML-/AI-based decision support systems, the data scientist, and the critical care practitioner.

One of the reasons why ML-based systems need to be interpretable in the critical care domain is closer to the domain itself. It has to do with the idea that interpretability is required specifically when there is incompleteness in the formulation of a problem [5], and such an incompleteness may be the result of a mismatch between the ML-modeling objectives and the critical care goals. The latter are mediated by clinical guidelines. This is not necessarily a bottleneck, especially if we see the achievement of interpretability as the key to make model performance and ICU guidelines compliance compatible. Clinical guidelines at the point of care are nothing but prior knowledge-based procedural decision-making pipelines. The idea of ML/AI data analytics is increasingly being referred to as “analytical pipeline” and, in a related approach, automated machine learning, or AutoML [10], which even addresses the possibility of automated pipeline

optimization. In turn, in the case of medical applications, this is akin to the more time-honored concept of medical decision support systems (MDSS), which have made inroads at least in specific domains [11, 12] and whose barriers to adoption have for long been subject of research [13]. The similarity of ML-based data analytics and medical guidelines conveyed in the concept of pipeline should be the key to stitch the former to the reality of the point-of-care. Note that, in the discussion about the weak levels of MDSS real adoption, it has been argued [13] this might be due to adequate “explanations [not being] given for the system’s diagnosis,” or “the form of explanation [not being] satisfactory for the physicians using the system.”

Where at the ICU Should We Apply AI and ML?

Critical care deals with the seriously or critically ill patients who have, are at risk of, or are recovering from life-threatening conditions. It includes providing life support, invasive monitoring techniques, resuscitation, and end-of-life care. Therefore, most of these patients are both pharmacologically and technologically dependent on the life-sustaining devices that surround them (infusion pumps, mechanical ventilators, external kidneys, catheters, and so on).

Patients are admitted into the ICU if their medical needs are greater than what the general ward can offer. Indications for ICU admission include severe hypertension/hypotension, sepsis, certain cardiac arrhythmias, major trauma, major burns, or brain damage, among others. Other ICU needs include airway or ventilator support due to respiratory compromise, such as severe cases of the COVID-19 pandemic happening at the time of writing this chapter. Patients may also be admitted into the ICU after a major surgery. In summary, critical care deals with the management of organ dysfunction regardless of its source whose cumulative effect is termed as multiple organ failure or multiple organ dysfunction syndrome. Intensive care specialists are also in charge of the “chain of survival” and, thus, apply cardiopulmonary

resuscitation (CPR), defibrillation, and, if necessary, advanced life support (ALS) as stated above. For this reason, beyond managing organ dysfunction, one of the most important needs in critical care is assessing whether a patient will suffer a sudden deterioration that may require resuscitation (i.e., “code blue”). The code blue is typically the result of failed interventions/treatments that eventually result in cardiac arrest and, finally, death.

In developed countries, sepsis (an infection causing organ dysfunction) is the most common immediate cause of death in hospitals [14], and despite decades of research for more effective cures, it affects between 47 and 50 million people every year [15], out of which at least 11 million die (1 death every 2.8 s, on average). Nevertheless, these terrifying numbers fail to convey the magnitude of the effect that the incidence of the new COVID-19 pandemic has generated. Deaths by COVID-19 are cases of viral sepsis resulting from an infection by SARS-CoV-2, which has arisen abruptly while the traditional bacterial sepsis remains in steady state, without real ups and downs in yearly incidence.

Over the last decade, over one hundred Phase II and Phase III clinical trials have taken place including several hundreds of thousands of patients on protocols to modify the systemic inflammatory response in sepsis [16]. Remarkably, none of these studies has resulted in new treatments. The challenges for overcoming sepsis are:

1. Incomplete understanding of the burden of sepsis
2. Absence of national and international guidelines for treatment and follow-up
3. Unacceptable long time to diagnosis
4. Lack of patient trajectory assessment systems for a personalized medicine approach to treatment

The diagnosis of sepsis is only confirmed when the patient is already in organ dysfunction, meaning that there are no established criteria for its early and accurate diagnosis.

Besides sepsis, acute lung injury (ALI) or mild acute respiratory distress syndrome (ARDS) is a diffuse heterogeneous lung injury characterized

by hypoxemia, non-cardiogenic pulmonary edema, low lung compliance, and widespread capillary leakage. Although it can be triggered by a respiratory infection (such as COVID-19 or pneumonia), it is also a result of sepsis and significant trauma. The incidence of this has significantly increased due to the incidence of SARS-CoV-2 infections resulting in the collapse of healthcare and intensive care units worldwide. The main guidelines for managing ALI include mechanical ventilation, pronation, and, in the most severe cases, induced coma.

Mechanical ventilation is not exempt of risks either. It is often associated with major complications. If it can be proven that mechanical ventilation caused an ALI, it is termed ventilator-induced lung injury (VILI). If the mechanical ventilator cause cannot be proven, then it is termed ventilatory-associated lung injury (VALI). VALI is the most common complication since it is virtually impossible to prove the cause of injury. One major causative factor of lung injury is the overstretching of the airways and alveoli (volutrauma). During mechanical ventilation, the flow of gas into the lung will take the path of least resistance. Areas of the lung that are collapsed (atelectasis) or filled with secretions will be under inflated, while those areas that are relatively normal will then become overinflated, which leads to distension and injury. The effect may be reduced by using smaller tidal volumes.

Another possible ventilator-associated lung injury is known as biotrauma. Biotrauma involves the lung suffering injury from any mediator of the inflammatory response or from sepsis. The mortality attributed to VILI has been significantly reduced in recent years, thanks to the protective measures applied to ventilated patients. Weaning from mechanical ventilation is the process of reducing ventilatory support, ultimately resulting in a patient breathing spontaneously and being extubated. The purpose is to assess the probability that mechanical ventilation can be successfully discontinued. Patients who wean successfully have less morbidity, mortality, and resource utilization than patients who require prolonged mechanical ventilation or the reinstitution of mechanical ventilation. The most successful

weaning strategies include a daily assessment of the patient's readiness to wean and the careful use of sedatives. Then, the daily screening of patients who are on mechanical ventilation with the aim of identifying those able to breathe spontaneously is, possibly, the best approach to reduce the duration of ventilatory support and sedation. Unnecessarily prolonging sedation and mechanical ventilation is directly associated with increased morbidity and mortality of critically ill patients as well as the length of ICU stay and a significant increase in costs. Therefore, a strategy to detect at an early stage and in real time the patients able to start weaning leads to a reduction in morbidity, mortality, length of ICU stay, and costs.

The integration of multilevel (i.e., sociodemographics, epidemiological, genetic, systemic, and cellular-level data) and multimodal data (such as patient records, monitoring data, lab tests, biomarkers, and so on) through ML-based solutions provides a valuable opportunity to further understand the mechanisms of organ dysfunction, its phenotyping, and assessing patient responses as well as the clinical response to interventions at the onset (or prior to, when possible) of organ dysfunction.

In this context, the current ML trend in critical care is to integrate multimodal/multilevel data with point-of-care devices for assessing specific protein and metabolic biomarkers to develop clinical decision support systems for diagnosis and early management of organ dysfunction. These solutions shall empower caregivers in the use of novel diagnostic tools, enabling the adoption of more effective management and personalized care by taking a systems-level approach for fine-tuning treatment and vital support. It is expected that these approaches will ultimately lead to a reduction in time to diagnose, days on vasopressor support, days on mechanical ventilation (i.e., early weaning), and, ultimately, mortality. These data modeling-centered integrated solutions aim at ensuring (i) the efficient prediction and assessment of organ dysfunction; (ii) the phenotyping and assessment of patients; and (iii) the timely delivery of treatment recommendations based on the assessment of the clinical trajectory of each patient.

What Sort of AI and ML Can We Apply at the ICU?

Providing an exhaustive summary of AI/ML applications on critical care is beyond the scope of this chapter. For that, the reader is referred to [12] and to a very recent review in [17]. Instead, a more focused review is provided. In a typical ML pipeline, data with available ethical committee approval and consent is collected, processed, pseudonymized, and included into a repository (often cloud-based) with a predefined data taxonomy. In the case of sepsis and ARDS/ALI, special emphasis is given to lactate measurements, C-reactive protein, blood pressure measurements, mechanical ventilation parameters, administration of antibiotics, interleukin blockers, crystalloids, and vasopressors. This basic information is enhanced with data from patient records, medical images, maneuvers, *in vivo/in vitro* tests, functional tests, monitoring data, other lab tests, and biomarkers that may include transcriptomics, proteomics, peptidomics, and metabolomics. In this regard, it is vital to carry out data collection in three main steps: collection and preparation, annotation, and harmonization. It is also important to perform a quality check on the data to assure completeness and absence of bias so that potential data repairs can be arranged.

In the case of sepsis, a major trend today in research is improving the accuracy of its diagnosis. The definition of sepsis was updated in 2016 [18] and advocated to the quick Sequential Organ Failure Assessment (qSOFA), which assesses blood pressure, respiratory rate, and mental status for the diagnosis of sepsis. A major criticism by the medical community of this score lays in its low specificity. For this reason, different research teams are trying to enhance this scale through the addition of bedside parameters (e.g., biomarker data) for enhancing this diagnosis criterion. These enhanced versions of the qSOFA scale are evaluated in the context of all available data at hospital admission through standard ML techniques such as multivariate logistic regression, relevance vector machines (RVM), support vector machines (SVM), shallow neural networks, or random forests (RF), to name a few, taking

the diagnosis of sepsis confirmed through hemocultures as main outcome.

Another key aspect in clinical research is obtaining a set of baseline phenotypes and patient trajectories in the ICU through multivariate analysis techniques such as principal component analysis (PCA), factor analysis (FA), and probabilistic clustering (PC). For example, [19] defines four different phenotypes for sepsis: patients with low vasopressor titration, patients with chronic conditions and lower renal function, patients with high inflammation and pulmonary dysfunction, and patients with liver dysfunction and septic shock through consensus k-means clustering. Another study [20] defines four phenotypes related to predicting ICU outcomes: patients requiring mechanical ventilation support, patients with high organ dysfunction, patients with high severity, and hepatic dysfunction. With the aim of predicting organ dysfunction before its onset, these phenotypes are now being improved through the addition of different clinical traits and biomarkers that become altered before organ dysfunction is detected at a systemic level. Moreover, current initiatives aim at further improving these phenotypes through the application of a generalization of the FA method with deep autoencoders (DAE) to assess the strength of associations between variables and their importance within each patient phenotype.

A common trend between these initiatives is that they all pave the way to study patient trajectories in the ICU. Patient trajectory assessment includes studying the prevalence of each phenotype as well as their impact on other clinical outcomes such as long-term survival (i.e., 100-day survival rate), pressor resistance, and days on organ support. Deep reinforcement learning (DRL) has also become an important line of research for assessing the continuum of organ dysfunction in sepsis. For example, [21] proposed a continuous state-space model for the management of sepsis in a twist that goes beyond the more traditional development and use of discriminative classifiers. Other studies have used Bayesian networks and Random Forests [22] for assessing patient trajectories of septic and septic shock patients in the acute phase. It is also expected

that with the inclusion of biomarker data from the complement cascade, platelet degranulation, acute inflammation response, negative regulation of endopeptidase activity, and blood coagulation, it is possible to further assess the continuum of organ dysfunction with high accuracy through the development of comprehensive, interpretable, and mathematically rigorous graphical model embeddings through deep learning techniques such as DRL and standard ML techniques such as conditional independence maps and generative kernels [23]. These techniques are not only expected to improve the accuracy in diagnosis and trajectory prediction and, in particular, long-term survival but also should set the basis for the personalized treatment of organ dysfunction.

A Further Few Things About the Use of AI and ML in Medicine that Merit Discussion

Most of the discussion so far in this chapter has revolved around the critical care areas to which AI and ML can be applied and about the type of AI and ML models that can be more suited to the analysis of the data generated by specific critical care problems. The use of AI for critical care is also mediated by a few other practical challenges. After all, we should wonder why this type of approaches has not really permeated medical practice yet, despite their technological maturity.

The use of AI and ML methods to assist decision-making in the critical care domain is, for instance, also likely to concern practitioners in terms of legal boundaries and their implications. In Europe, this directly relates to the recent implementation of the European Union directive for General Data Protection Regulation (GDPR), which mandates a “right to explanation” of all decisions made by “automated or artificially intelligent algorithmic systems” [24]. Article 13 of this directive indicates that a designated “data controller” is legally bound to provide requesting citizens with “meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing [automated decision making, as described in its Article 22] for

the data subject.” As stated in [25], at least part of the problem has to do with data pseudonymization and anonymization. If proper anonymization is in place, data sharing and analysis should, in principle, not collide with legal privacy issues. The paradox is that ML, particularly in the form of deep learning, has been shown to be able to challenge medical data anonymization procedures [26]. Note also that there is a subtle and somehow related problem: the fact that ML and related methods often require substantial databases in order to extract useful and generalizable knowledge; gathering such databases faces non-trivial challenges, because it would require linking databases across clinical centers and countries (i.e., it would require the use of multi-center, international databases). As discussed in [27], this might boil down to another issue also brought up in [25]: the matter of data property. Who holds the data ownership at the ICU? Is it the patients? The professionals who have collected these data or the medical institutions they belong to? The external companies developing medical software used at the point-of-care? [28] If anonymized data are going to be used for ML-based analysis, does it require patients’ informed consent in all cases? The problem here, clearly expressed in [27], is the heterogeneity with which ethics committees at clinical centers deal with the requirement for consent for secondary uses of health data (semi-automated analysis in this case) provided certain conditions are met. And we are not talking only about the whim of particular ethics committees but also about the fact that these committees might be bounded to comply with heterogeneous local, national (such as the US HIPAA), or even transnational legislations (like GDPR). Such heterogeneity precludes the existence of high-quality databases for ML analysis, and the way forward is unclear but would entail either legislative harmonization or, perhaps in a more feasible manner, achievable in a shorter term, negotiation of codes of conduct related to data handling between professional bodies.

Note that GDPR legislation, in any case, is only partially about data rights as such and requires something very specific from the people implementing and using ML systems: the

aforementioned “right to explanation”: this leads back to the interpretability and explainability issues discussed in previous sections.

Most of the discussion so far in this chapter has revolved around the critical care areas to which AI and ML can be applied and about the type of AI and ML models that can be more suited to the analysis of the data generated by specific critical care problems. There is a further underlying problem that is not often broached [29] in relation to the previous: the use of open-source software (OSS) in the implementation of ML-based analytical pipelines and the potential benefits and risks of endowing critical care practitioners with the skills required to use such software.

Over the last few years, there has been a very swift transition from the predominance of proprietary software for the development of ML-based data analysis to the expansion of the use of non-proprietary software for such development, not only in the academic domain but also in applied domains traditionally reluctant to the use of OSS. From this, we can interpret that the advantages of OSS have proved to be more than those of the proprietary counterparts and that their inconveniences and barriers-to-use have been limited, if not vanished. It can be argued that the medical domain in general has been more impervious to these changes but also that [30] software vendors have failed to provide stable technical partnerships with the clinical domain in terms of medical informatics. Proprietary data management and processing also limit the reach of standardization of proprietary technology, therefore limiting as well the ability to take full advantage of patients’ population data analytics. OSS eliminates licensing costs, promotes compatibility, and allows customizing the software tools to the medical domain and its requirements [31], encouraging collaborative innovation and shortening software development cycles [32]. Why should not the critical care domain take full advantage of OSS for the development of ML-based MDSS? This coalesces with the recent research push for the development of automated ML or AutoML tools in medicine [33], despite existing and nontrivial barriers related to potential unreliability and unwanted variability of

the data, potential inadequacy to large-scale datasets, and also the lack of interpretability and explainability of many ML methods that we have earlier discussed in this chapter [10].

Conclusions

Should intensive care physicians care about AI and ML in their application to their domain of expertise? Arguably not. Should a data analyst specialized in medicine and clinical applications care about the application of AI and ML in critical care? Arguably yes. This asymmetry is a sobering reminder that there is plenty to do before these new technologies become a full-time companion for medical decision-making in critical care. We have provided in this chapter a general overview of how AI and ML can contribute in different problems of the domain, but, perhaps more importantly, we have highlighted the characteristics of many of the challenges yet to overcome.

As stated from the onset, the data wealth provided by the necessary use of digital equipment at the point of care, coalescing with the use of omics data, makes the intensive ward the perfect environment to take advantage of advanced data analytics. As we have argued, only by achieving a seamless integration with medical guidelines, by taking on board the human expertise, and by facilitating personalized medical treatment will AI and ML deliver on the great expectations that have been placed upon them.

References

1. Cosgriff CV, Celi LA, Stone DJ. Critical care, critical data. *Biomed Eng Comput Biol.* 2019;10:1179597219856564.
2. Ravi D, Wong C, Deligianni F, Berthelot M, Andreu Pérez J, Lo B, Yang GZ. Deep learning for health informatics. *IEEE J Biomed Health.* 2017;21(1):4–21.
3. Cabitzza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *JAMA.* 2017;318(6):517–8.
4. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J Clin Epidemiol.* 1996;49(11):1225–31.

5. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. 2017. arXiv preprint. arXiv:1702.08608.
6. Bacciu D, Lisboa PJ, Martín JD, Stoean R, Vellido A. Bioinformatics and medicine in the era of deep learning. In: Proceedings of the 26th European symposium on artificial neural networks, computational intelligence and machine learning (ESANN 2018), Bruges, 2018. p. 345–54.
7. Vellido A, Martín JD, Rossi F, Lisboa PJG. Seeing is believing: the importance of visualization in real-world machine learning applications. In: Proceedings of the 19th European symposium on artificial neural networks (ESANN), 2011. p. 219–26.
8. Bhanot G, Biehl M, Villmann T, Zühlke D. Biomedical data analysis in translational research: integration of expert knowledge and interpretable models. In: Proceedings of the 25th European symposium on artificial neural networks, computational intelligence and machine learning (ESANN), 2017. p. 177–86.
9. Vellido A. The importance of interpretability and visualization in Machine Learning for applications in medicine and health care. *Neural Comput Appl*. ePub ahead of press. <https://doi.org/10.1007/s00521-019-04051-w>.
10. Waring J, Lindvall C, Umeton R. Automated machine learning: review of the state-of-the-art and opportunities for healthcare. *Artif Intell Med*. 2020;104:101822.
11. Saifdar S, Zafar S, Zafar N, Khan NF. Machine learning based decision support systems (DSS) for heart disease diagnosis: a review. *Artif Intell Rev*. 2017;50(4):597–623.
12. Vellido A, Ribas V, Morales C, Ruiz-Samartín A, Ruiz-Rodríguez JC. Machine learning for critical care: state-of-the-art and a sepsis case study. *Biomed Eng Online*. 2018;17(S1):135.
13. Dreiseitl S, Binder M. Do physicians value decision support? A look at the effect of decision support systems on physician opinion. *Artif Intell Med*. 2005;33(1):25–30.
14. Rhee C, Jones TM, Hamad Y, Pande A, Varon J, O'Brien C, Anderson DJ, Warren DK, Dantes RB, Epstein L, Klompaas M. Prevalence, underlying causes, and preventability of sepsis-associated mortality in US acute care hospitals. *JAMA Netw Open*. 2019;2(2):e187571.
15. <https://www.global-sepsis-alliance.org/sepsis>
16. Marshall JC. Why have clinical trials in sepsis failed? *Trends Mol Med*. 2014;20(4):195–203.
17. Nguyen D, Ngo B, van Sonnenberg E. AI in the intensive care unit: up-to-date review. *J Intensive Care Med*. 2020. ePub ahead of publication. <https://doi.org/10.1177/0885066620956620>.
18. Seymour CW, Liu VX, Iwashyna TJ, Brunkhorst FM, Rea TD, Scherag A, Rubenfeld G, Kahn JM, Shankar-Hari M, Singer M, Deutschman CS. Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*. 2016;315(8):762–74.
19. Seymour CW, Kennedy JN, Wang S, Chang CC, Elliott CF, Xu Z, Berry S, Clermont G, Cooper G, Gomez H, Huang DT. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. *JAMA*. 2019;321(20):2003–17.
20. Ribas VJ, Vellido A, Ruiz-Rodríguez JC, Rello J. Severe sepsis mortality prediction with logistic regression over latent factors. *Expert Syst Appl*. 2012;39(2):1937–43.
21. Raghu A, Komorowski M, Celi LA, Szolovits P, Ghassemi M. Continuous state-space models for optimal sepsis treatment—a deep reinforcement learning approach. arXiv preprint arXiv:1705.08422. 2017 May 23.
22. Aushev A, Ribas Ripoll V, Vellido A, Aletti F, Bollen Pinto B, Bendjelid K, Herpain A, Hendrik Post E, Romay Medina E, Ferrer R, Baselli G. Feature selection for the accurate prediction of septic and cardio-genic shock ICU mortality in the acute phase. *PLoS One*. 2018;13(11):e0199089.
23. Ripoll VJ, Vellido A, Romero E, Ruiz-Rodríguez JC. Sepsis mortality prediction with the quotient basis kernel. *Artif Intell Med*. 2014;61(1):45–52.
24. Goodman B, Flaxman S. European Union regulations on algorithmic decision making and a “right to explanation”. *AI Mag*. 2017;38(3):50–57.
25. Reiz AN, de la Hoz MA, García MS. Big data analysis and machine learning in intensive care units. *Med Intensiva (English Edition)*. 2019;43(7):416–26.
26. Yoon J, Drumright LN, Van Der Schaar M. Anonymization through data synthesis using generative adversarial networks (ADS-GAN). *IEEE J Biomed Health Inform*. 2020;24(8):2378–88.
27. McLennan S, Shaw D, Celi LA. The challenge of local consent requirements for global critical care databases. *Intensive Care Med*. 2019;45(2):246–8.
28. Tanner A. Our bodies, our data: how companies make billions selling our medical records. Beacon Press; 2017.
29. Hueso M, de Haro L, Calabia J, Dal-Ré R, Tebé C, Gibert K, Cruzado JM, Vellido A. Leveraging data science for a personalized haemodialysis. *Kidney Dis*. 2020. ePub ahead of print. <https://doi.org/10.1159/000507291>.
30. Caban JJ, Joshi A, Nagy P. Rapid development of medical imaging tools with open-source libraries. *J Digit Imaging*. 2007;20(Suppl 1):83–93.
31. Karopka T, Schmuhl H, Demski H. Free/libre open source software in health care: a review. *Healthc Inform Res*. 2014;20(1):11–22.
32. Aminpour F, Sadoughi F, Ahamdi M. Utilization of open source electronic health record around the world: a systematic review. *J Res Med Sci*. 2014;19(1):57–64.
33. Alaa AM, van der Schaar M. Autoprognosis: automated clinical prognostic modeling via Bayesian optimization with structured kernel learning. arXiv preprint arXiv:1802.07207. 2018.



Artificial Intelligence in Medicine (AIM) for Cardiac Arrest

107

Hisaki Makimoto

Contents

Introduction	1480
Cause of Cardiac Arrest: Ventricular Arrhythmias	1480
Baseline Cardiovascular Diseases Leading to Ventricular Arrhythmia	1481
From Ventricular Arrhythmias to Cardiac Arrest	1481
The Current Clinical Treatment to Prevent Cardiac Arrest	1481
The Possibility and Advantage to Introduce AI Technology	1482
Overview of Researches to Prevent Cardiac Arrest Utilizing AI	1482
Early Recognition/Detection of High-Risk Patients	1482
Active Intervention and Follow-up	1483
Continuous Monitoring and Subsequent Interventions	1483
Discussion	1484
References	1484

Abstract

Cardiac arrest is one of the major causes of death worldwide. There is a significant clinical and economical need to prevent cardiac arrest.

The risk of cardiac arrest increases with age and comorbidities such as hypertension or other cardiovascular diseases. As the average life expectancy is increasing, the number of people

predisposed to cardiac arrest is increasing globally. Moreover, cardiac arrest can also occur in the young, healthy population and may result in sudden cardiac death due to hereditary diseases. The clinical and societal needs to prevent cardiac arrest continue to grow.

The recent development of artificial intelligence (AI) has shown potential for improving clinical techniques in the field of cardiovascular medicine. There are three important phases of interventions to prevent cardiac arrest: (1) early recognition of high-risk populations, (2) active interventions and follow-up according to individual estimated sudden cardiac arrest risks, and

H. Makimoto (✉)
Arrhythmia Service, Division of Cardiology, Pulmonology and Vascular Medicine, Faculty of Medicine, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany

(3) continuous monitoring in daily life and subsequent interventions. AI technology has been introduced in each phase. In this chapter, we discuss the recent achievements of AI in medicine (AIM) related to cardiac arrest and discuss its potential future uses.

Keywords

Sudden cardiac death · Heart failure · Ventricular tachycardia · Wearable devices

Introduction

Cardiac arrest is defined as the cessation of heart function resulting in ineffective pumping of the heart and possibly leading to death. A typical cause of cardiac arrest is acute myocardial infarction. Chest pain, faintness, and nausea are possible symptoms that can occur shortly before cardiac arrest.

Sudden cardiac arrest (SCA) is an abrupt form of cardiac arrest. In developed countries, SCA is considered a major cause of death, accounting for 15% of all deaths [1]. Although acute myocardial infarction is well known for occurring during hard exercises or when the body is rapidly and intensely cooled, for example, in winter, there are many cases of SCA in which the normal heart beats suddenly cease without any obvious pathology. SCA is a life-threatening condition and can be difficult to predict. Therefore, it is a major clinical concern.

The most typical and direct cause of SCA is a ventricular arrhythmia. Risk factors of ventricular arrhythmias include age and baseline cardiovascular diseases, particularly ischemic heart diseases such as angina pectoris, myocardial infarction, heart failure, and various cardiomyopathies [2]. Some hereditary arrhythmic diseases have also been reported to cause SCA in relatively young patients. Bradycardia, in which the heart rate is slower than expected, such as in atrioventricular block, can also cause SCA.

The current standard medical treatments for high-risk patients who are exposed to SCA are implantations of medical devices such as implantable cardioverter-defibrillator (ICD) or

pacemakers [2, 3]. These devices can detect pre-defined heart rhythms that may be harmful to the patient. This detection leads to heart stimulation or shock delivery through electrical signals to retain an appropriate rhythm. The intracardiac electrodes are placed on the atrial or ventricular myocardium, and the generator is placed subcutaneously or submuscularly in the pectoral region of the anterior chest wall. The main function of the ICD is to detect ventricular arrhythmias and subsequently terminate them using stimulation or electrical shocks. Pacemakers stimulate and maintain heart beats if the heart rate is slower than predefined rates.

Currently, these medical devices are the most effective preventions against SCA. However, there are shortcomings of these devices, including higher risks of infections locally and intracardially and the risk of inadequate or inappropriate intervention of the devices (e.g., electrical shocks due to erroneous detection of abnormal heart rhythms) [4]. Thus, the necessity and suitability of these devices must be carefully examined before implantation, considering each patient's risk of SCA. Unfortunately, it is difficult or nearly impossible to accurately predict if the concerned patient will experience SCA. Artificial intelligence (AI) technology may supplement this decision-making process with its high performance in memory, recognition, and analysis. Some studies have reported the use of AI to prevent SCA more efficiently. In this chapter, we discuss some achievements of SCA prediction and prevention using AI and discuss further possibilities for the use of AI in medicine (AIM).

Cause of Cardiac Arrest: Ventricular Arrhythmias

The typical cause of cardiac arrest is a ventricular arrhythmia, most of which are ventricular tachycardia or ventricular fibrillation. Ventricular arrhythmias lead to ventricle trembles, resulting in a loss of the pump function of the ventricle, leading to circulatory collapse and cardiac arrest. Ventricular bradycardia can also result in circulatory collapse and cardiac arrest if the heart rate is extremely slow.

Ventricular tachycardia is defined as rapid and regular heartbeats (≥ 100 bpm) generated in the ventricular myocardium, contrary to normal heartbeats generated in the right atrium of the heart (at the sinus node). Ventricular fibrillation is a faster and finer trembling of the ventricular myocardia.

Ventricular bradycardia is defined as a heartbeat ≤ 60 bpm, and severe cases of bradycardia (heartbeat ≤ 30 bpm) can lead to circulatory collapse.

Baseline Cardiovascular Diseases Leading to Ventricular Arrhythmia

These ventricular arrhythmias rarely occur in healthy young adults who do not have an underlying cardiovascular disease. Ventricular tachycardia and fibrillation can be more frequently observed in elderly patients with a history of myocardial infarction or heart failure [5]. Various cardiomyopathies and inherited cardiovascular diseases have also been reported to be risk factors for these ventricular arrhythmias [2].

In the acute phase of myocardial infarction, myocardia undergo necrosis due to an abrupt disruption of the oxygen supply [6]. A typical cause of this is the obstruction of the coronary artery that supplies oxygen to the myocardium by a thrombus generated from the rupture of atherosclerotic plaques. Patients with hypertension, diabetes mellitus, dyslipidemia, renal dysfunction, and atrial fibrillation are considered to be high risk, with a predisposition for acute myocardial infarction [6].

Heart failure is broadly considered to be a condition in which cardiac dysfunction causes symptoms such as dyspnea. When the pump function of the ventricles is severely reduced, the patient may develop heart failure. Heart failure can result from primary myocardial damage due to dilated cardiomyopathy or hypertrophic cardiomyopathy as well as from secondary myocardial damage due to myocardial infarction or subsequent ischemic cardiomyopathy.

Some hereditary arrhythmic disorders are frequently seen in relatively young patients who experience ventricular tachycardia or fibrillation. These hereditary disorders involve myocardial ion channels that contribute to myocardial action

potentials and include Brugada syndrome, long QT syndrome, and catecholaminergic polymorphic ventricular tachycardia.

From Ventricular Arrhythmias to Cardiac Arrest

Patients with a prior history of myocardial infarction or heart failure have injured myocardium that no longer functions normally. These injured myocardial tissues are termed “arrhythmogenic substrates” and can generate ventricular tachycardia or fibrillation. In the case of hereditary arrhythmic disorders, myocardial ion channels harbor genetic dysfunctions, and these abnormalities precipitate ventricular arrhythmias.

When the heart loses its pump function due to ventricular arrhythmia, the blood pressure falls, resulting in circulatory collapse. In most cases, patients lose consciousness 10 s after the ventricular arrhythmia.

The Current Clinical Treatment to Prevent Cardiac Arrest

The current standard treatment for SCA due to ventricular arrhythmias is implantation of an implantable cardioverter-defibrillator (ICD). ICDs can detect the occurrence of the ventricular arrhythmias based on predefined settings and terminate arrhythmias through electrical stimulation and shocks. ICD implantation is the standard treatment to prevent SCA due to ventricular arrhythmias [2]. There are two basic indications for ICD implantation. Primary prophylactic ICD implantation is indicated in patients with highly reduced left ventricular function who have had no prior sustained ventricular arrhythmias. Secondary prophylactic ICD implantation is indicated in patients with a history of ventricular arrhythmia events. The latter patients have a naturally higher incidence of subsequent ICD interventions (stimulation/shocks) compared to the former patients.

For patients with bradycardia, the implantation of a pacemaker can maintain the heart rate and contribute to the prevention of cardiac arrest through direct myocardial stimulation [3]. The

necessity of pacemaker implantation should be determined by the patient's symptoms and the specific type of heart block.

The implantation of these devices also has disadvantages. In patients with implanted devices, especially for primary prevention, ventricular arrhythmia does not always occur following implantation. It is difficult to identify patients who will benefit from implantation prior to the procedure as future ventricular arrhythmias leading to cardiac arrest are difficult to predict in these patients. The implantation of these devices may also accompany long-term risks of complications [4]. The device itself causes discomfort and, furthermore, may give inappropriate shocks when the device misdiagnoses arrhythmias. Moreover, stimulations or shocks to terminate ventricular tachycardias have been reported to predict higher mortality, whether the interventions were appropriate or not, although ICDs effectively terminated life-threatening ventricular arrhythmias [7, 8].

To overcome these shortcomings of implantable devices, a more reliable and efficient method of estimating the risk of future ventricular arrhythmias is needed to determine the necessity of ICD implantation.

The Possibility and Advantage to Introduce AI Technology

AI technology may be a more efficient way to predict and prevent cardiac arrest. In theory, AI is not only able to gain knowledge via a learning process, similar to humans, but also to recognize information in a more efficient manner than human physicians. Furthermore, AI can manage large data simultaneously at high speed due to the development of computational technologies in the last decade. AI technology has begun to be utilized to prevent cardiac arrest.

Overview of Researches to Prevent Cardiac Arrest Utilizing AI

Machine learning has been a mainstay technique to build AI for cardiovascular data. In recent years, more researchers have adopted deep

learning models instead of conventional algorithm-based models.

Electrocardiogram (ECG) data are most commonly used as input data as the ECG is one of the major examinations used in cardiovascular medicine. Other medical data including blood tests, blood pressure measurements, and video or photographs of the face have also been used to build the AI for cardiovascular medicine. Simple data are preferred as input data, as the priority for preventing cardiac arrest is to determine or predict lethal ventricular arrhythmias as soon as possible, rather than to predict an exact diagnosis. For example, the ECGs used in some reports were recorded using a single lead, in contrast to the 12 leads used in the clinic. The quantity of data used for AI building ranges from that of several hundreds to tens of thousands of patients.

The target cardiac diseases to be identified by AI include myocardial infarction, cardiac dysfunction (heart failure), and arrhythmias, all of which are related to cardiac arrest. AI has been used to detect anomalies or diagnose these cardiac diseases.

For the prevention of cardiac arrest, there are three important phases of intervention:

1. Early recognition/detection of high-risk patients
2. Active interventions and follow-up according to individual estimated SCA risks
3. Continuous monitoring in daily life and subsequent interventions

Here, we present the achievements of AI at each phase to date.

Early Recognition/Detection of High-Risk Patients

It is important to identify high-risk patients as early and precisely as possible to prevent cardiac arrest. A history of sustained ventricular arrhythmias (i.e., previous aborted sudden cardiac arrest) and highly reduced left ventricular function are the most reliable risk predictors for cardiac arrest. Left ventricular function is determined comprehensively using echocardiography and cardiovascular magnetic resonance (CMR) [2].

Researchers in the United States have reported that patients with highly reduced left ventricular function can be detected with ECG data only by AI built using deep learning [9]. The researchers reported that a convolutional neural network (CNN) with nine layers was trained with ECGs from 44,959 patients and tested with ECGs from 52,870 patients. This CNN could distinguish patients with severe left ventricular dysfunction using only raw ECG data with an accuracy of 85% (area under the curve = 0.93). Interestingly, some patients classified as severe left ventricular dysfunction by the CNN but whose cardiac function had not deteriorated significantly according to echocardiography developed a severe decline in cardiac function upon follow-up. Therefore, this CNN identified the current high-risk patients as well as the potential high-risk patients using only ECG data.

There have been attempts to build AIs using video capturing with smartphones. Researchers in China reported an AI program built via deep learning to predict blood pressure using videos of the human face obtained with a smartphone [10]. Based on transdermal optical imaging from 2-min-long videos, the blood pressure could be predicted through the multilayer perceptron in the error range of 5 ± 8 mmHg. Higher blood pressure is correlated with a higher incidence of acute myocardial infarction or heart failure leading to cardiac arrest. This program may result in the very early discovery of patients with the potential to be at high risk for cardiac arrest.

The smartphone videos were also used for the detection of arrhythmias [11]. A CNN that could detect atrial fibrillation from video data with an accuracy of more than 95% both in sensitivity and specificity (AUC = 0.99) has been reported. These reports suggest that AI can currently recognize information that human physicians cannot.

Active Intervention and Follow-up

The risk of cardiac arrest often increases if patients are affected by certain cardiovascular diseases, such as myocardial infarction. In these patient populations, medical interventions should address long-term care and subsequent possible

cardiac arrest risks in addition to providing direct medical therapy for cardiovascular diseases. AI technology has started to be used to efficiently identify these patients.

Risk scoring systems are currently used to estimate the mortality or rehospitalization risk of patients with cardiovascular diseases. It has been reported that a machine learning algorithm built using a large amount of electronic health record (EHR) data can predict mortality and rehospitalization more precisely than the current risk scoring system [12]. Utilizing the EHR data in combination with other clinical data, the mortality of patients with heart failure could be predicted more precisely than with the Seattle Heart Failure Model (SHFM), which is the clinical index used to determine risks for patients with heart failure. Another study reported that AI built on a machine learning algorithm can use EHR data to precisely predict the onset of heart failure in a span of 12–18 months [13].

Torsade de pointes, a specific type of lethal ventricular tachycardia, can be induced by the administration of medications, which was predicted during a simulation using machine learning [14]. This study reported that a computer model was built based on the electrophysiological characteristics of cardiomyocytes and their reaction to drugs. This computer program was able to predict the risk of an order-made medication according to the electrophysiological properties of the patient.

Continuous Monitoring and Subsequent Interventions

Patients considered to be at high-risk for cardiac arrest are monitored and protected everyday by implanted devices (ICDs or pacemakers). Noninvasive monitoring devices, such as handheld ECG equipment for single-lead, 3-lead, or 6-lead ECG, are also used. The algorithms of these devices, built using machine learning, can detect atrial fibrillation (AF), including subclinical AF. Subclinical AF may lead to a higher risk of stroke and myocardial infarction if not detected. The diagnosis of subclinical AF enables the initiation of anticoagulant therapy earlier to prevent subsequent attacks.

In the iREAD Study [15], the ability of single-lead handheld ECG equipment (KardiaMobile Cardiac Monitor (KMCM)) to detect AF was verified. This study reported that the KMCM detected AF as accurately as the physicians who interpreted single-lead and 12-lead ECGs (KMCM sensitivity, 100%; specificity, 89.2%).

Recently, more and more wearable devices with cardiac monitoring functions are being used by patients with cardiovascular diseases as well as the healthy population. Most of these devices work in conjunction with smartphones, where applications are developed to obtain and manage the healthcare data from the wearable devices. Some devices are equipped with algorithms built through machine learning, one of which was tested for its detection capability of AF [16]. The Apple Heart Study included 419,297 owners of Apple Watches and iPhones, and their pulse analysis was verified. According to the investigation of 86 cases in which an irregular pulse notification on the apple watch and the ECG monitoring could be collected simultaneously, the positive predictive value for the AF detection based on the irregular pulse notification of the apple watches was 84%.

Data from implanted medical devices such as ICDs or pacemakers can also be continuously gathered and analyzed in the background. This data has been used for the early detection of heart failure [17]. It has also been suggested that the mortality of patients with heart failure can be predicted more precisely by combining intra-device variables with other medical data [18].

Discussion

Cardiac arrest remains as a major cause of death worldwide. AI technology has great potential and is expected to improve the methods of cardiac arrest prevention. Healthcare data and medical data are mainly objective and easy to digitalize, which are appropriate for building an AI algorithm. Input data will continue to accumulate with the widespread use of wearable devices and smartphones. It is expected that the quantity of input data will increase due to the era of big data,

enhancing and improving AI constructions. AI can handle enormous amounts of data, such as genetic information, suggesting the possibility of applying it to the treatment of hereditary diseases and to individualized medical therapies. Researchers have reported that the onset of asthma can be predicted based on genetic data (a single-nucleoside polymorphism (SNP)) using machine learning [19]. This strategy may also be applied to cardiovascular diseases, for example, hereditary arrhythmia syndromes.

However, there remain problems and possible weakness in using AI technology in cardiovascular medicine. First, AI programs could not provide reasoning for their conclusions, making it difficult to determine if the AI has correctly learned by machine learning or deep learning methods. As suggested in Nature (2019) [20], AI algorithms may be easily deceived by data corruption that human physicians cannot perceive. AI technology is incomplete and must be carefully introduced in the medical field to ensure patient safety. Whether AI can be relied upon is dependent on technical development as well as political and legal factors. Furthermore, to build a better AI program, the quality of input data is crucial. Medical data are typically collected at hospitals by certified medical staff; however, healthcare data obtained by wearable devices are not verified by trained individuals. To build an AI, the quality of these input data must be well considered.

Cardiac arrest is a life-threatening condition with a significant effect on quality of life and healthcare costs. There is a need for a more efficient and simple way to predict and prevent cardiac arrest, and for this aim, AI technology may play a key role. It is strongly expected that systems or devices with AI technology will soon be put into practical use as an aid to human physicians.

References

1. Zheng ZJ, Croft JB, Giles WH, Mensah GA. Sudden cardiac death in the United States, 1989 to 1998. *Circulation*. 2001;104(18):2158–63. <https://doi.org/10.1161/hc4301.098254>.

2. Priori SG, Blomström-Lundqvist C, Mazzanti A, Blom N, Borggrefe M, Camm J, Elliott PM, Fitzsimons D, Hatala R, Hindricks G, Kirchhof P, Kjeldsen K, Kuck KH, Hernandez-Madrid A, Nikolaou N, Norekval TM, Spaulding C, Van Veldhuisen DJ, ESC Scientific Document Group. 2015 ESC guidelines for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death: The Task Force for the Management of Patients with Ventricular Arrhythmias and the Prevention of Sudden Cardiac Death of the European Society of Cardiology (ESC). Endorsed by: Association for European Paediatric and Congenital Cardiology (AEPC). *Eur Heart J.* 2015;36(41):2793–867. <https://doi.org/10.1093/euroheartj/ehv316>.
3. Brignole M, Auricchio A, Baron-Esquivias G, Bordachar P, Boriani G, Breithardt OA, Cleland J, Deharo JC, Delgado V, Elliott PM, Gorenek B, Israel CW, Leclercq C, Linde C, Mont L, Padeletti L, Sutton R, Vardas PE, ESC Committee for Practice Guidelines (CPG), Zamorano JL, Achenbach S, Baumgartner H, Bax JJ, Bueno H, Dean V, Deaton C, Erol C, Fagard R, Ferrari R, Hasdai D, Hoes AW, Kirchhof P, Knutti J, Kolp P, Lancellotti P, Linhart A, Nihoyannopoulos P, Piepoli MF, Ponikowski P, Sirnes PA, Tamargo JL, Tendera M, Torbicki A, Wijns W, Windecker S, Document Reviewers, Kirchhof P, Blomstrom-Lundqvist C, Badano LP, Aliyev F, Bänsch D, Baumgartner H, Bsata W, Buser P, Charron P, Daubert JC, Dobreaun D, Faerstrand S, Hasdai D, Hoes AW, Le Heuzey JY, Mavrakis H, McDonagh T, Merino JL, Nawar MM, Nielsen JC, Pieske B, Poposka L, Ruschitzka F, Tendera M, Van Gelder IC, Wilson CM. 2013 ESC guidelines on cardiac pacing and cardiac resynchronization therapy: the Task Force on cardiac pacing and resynchronization therapy of the European Society of Cardiology (ESC). Developed in collaboration with the European Heart Rhythm Association (EHRA). *Eur Heart J.* 2013;34(29):2281–329. <https://doi.org/10.1093/euroheartj/eht150>.
4. van der Heijden AC, Borleffs CJ, Buitenhuis MS, Thijssen J, van Rees JB, Cannegieter SC, Schalij MJ, van Erven L. The clinical course of patients with implantable cardioverter-defibrillators: extended experience on clinical outcome, device replacements, and device-related complications. *Heart Rhythm.* 2015;12(6):1169–76. <https://doi.org/10.1016/j.hrthm.2015.02.035>.
5. Khurshid S, Choi SH, Weng LC, Wang EY, Trinquart L, Benjamin EJ, Ellinor PT, Lubitz SA. Frequency of cardiac rhythm abnormalities in a half million adults. *Circ Arrhythm Electrophysiol.* 2018;11(7):e006273. <https://doi.org/10.1161/CIRCEP.118.006273>.
6. Thygesen K, Alpert JS, Jaffe AS, Chaitman BR, Bax JJ, Morrow DA, White HD, Executive Group on behalf of the Joint European Society of Cardiology (ESC)/American College of Cardiology (ACC)/American Heart Association (AHA)/World Heart Federation (WHF) Task Force for the Universal Definition of Myocardial Infarction. Fourth universal definition of myocardial infarction (2018). *J Am Coll Cardiol.* 2018;72(18):2231–64. <https://doi.org/10.1016/j.jacc.2018.08.1038>.
7. Daubert JP, Zareba W, Cannom DS, McNitt S, Rosero SZ, Wang P, Schuger C, Steinberg JS, Higgins SL, Wilber DJ, Klein H, Andrews ML, Hall WJ, Moss AJ, MADIT II Investigators. Inappropriate implantable cardioverter-defibrillator shocks in MADIT II: frequency, mechanisms, predictors, and survival impact. *J Am Coll Cardiol.* 2008;51(14):1357–65. <https://doi.org/10.1016/j.jacc.2007.09.073>.
8. van Rees JB, Borleffs CJ, de Bie MK, Stijnen T, van Erven L, Bax JJ, Schalij MJ. Inappropriate implantable cardioverter-defibrillator shocks: incidence, predictors, and impact on mortality. *J Am Coll Cardiol.* 2011;57(5):556–62. <https://doi.org/10.1016/j.jacc.2010.06.059>. PMID: 21272746.
9. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, Pellikka PA, Enriquez-Sarano M, Noseworthy PA, Munger TM, Asirvatham SJ, Scott CG, Carter RE, Friedman PA. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat Med.* 2019;25(1):70–4. <https://doi.org/10.1038/s41591-018-0240-2>.
10. Luo H, Yang D, Barszczuk A, Vempala N, Wei J, Wu SJ, Zheng PP, Fu G, Lee K, Feng ZP. Smartphone-based blood pressure measurement using transdermal optical imaging technology. *Circ Cardiovasc Imaging.* 2019;12(8):e008857. <https://doi.org/10.1161/CIRCIMAGING.119.008857>.
11. Yan BP, Lai WHS, Chan CKY, Au ACK, Freedman B, Poh YC, Poh MZ. High-throughput, contact-free detection of atrial fibrillation from video with deep learning. *JAMA Cardiol.* 2020;5(1):105–7. <https://doi.org/10.1001/jamacardio.2019.4004>.
12. Panahiazar M, Taslimitehrani V, Pereira N, Pathak J. Using EHRs and machine learning for heart failure survival analysis. *Stud Health Technol Inform.* 2015;216:40–4.
13. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. *J Am Med Inform Assoc.* 2017;24(2):361–70. <https://doi.org/10.1093/jamia/ocw112>.
14. Lancaster MC, Sobie EA. Improved prediction of drug-induced Torsades de Pointes through simulations of dynamics and machine learning algorithms. *Clin Pharmacol Ther.* 2016;100(4):371–9. <https://doi.org/10.1002/cpt.367>.
15. William AD, Kanbour M, Callahan T, Bhargava M, Varma N, Rickard J, Saliba W, Wolski K, Hussein A, Lindsay BD, Wazni OM, Tarakji KG. Assessing the accuracy of an automated atrial fibrillation detection algorithm using smartphone technology: the iREAD Study. *Heart Rhythm.* 2018;15(10):1561–5. <https://doi.org/10.1016/j.hrthm.2018.06.037>.

16. Perez MV, Mahaffey KW, Hedlin H, Rumsfeld JS, Garcia A, Ferris T, Balasubramanian V, Russo AM, Rajmane A, Cheung L, Hung G, Lee J, Kowey P, Talati N, Nag D, Gummidi Pundi SE, Beatty A, Hills MT, Desai S, Granger CB, Desai M, Turakhia MP, Apple Heart Study Investigators. Large-scale assessment of a smartwatch to identify atrial fibrillation. *N Engl J Med.* 2019;381(20):1909–17. <https://doi.org/10.1056/NEJMoa1901183>.
17. Conraads VM, Tavazzi L, Santini M, Oliva F, Gerritsse B, Yu CM, Cowie MR. Sensitivity and positive predictive value of implantable intrathoracic impedance monitoring as a predictor of heart failure hospitalizations: the SENSE-HF trial. *Eur Heart J.* 2011;32(18):2266–73. <https://doi.org/10.1093/eurheartj/ehr050>.
18. Manlucu J, Sharma V, Koehler J, Warman EN, Wells GA, Gula LJ, Yee R, Tang AS. Incremental value of implantable cardiac device diagnostic variables over clinical parameters to predict mortality in patients with mild to moderate heart failure. *J Am Heart Assoc.* 2019;8(14):e010998. <https://doi.org/10.1161/JAHA.118.010998>.
19. Gaudillo J, Rodriguez JJR, Nazareno A, Baltazar LR, Vilela J, Bulalacao R, Domingo M, Albia J. Machine learning approach to single nucleotide polymorphism-based asthma prediction. *PLoS One.* 2019;14(12): e0225574. <https://doi.org/10.1371/journal.pone.0225574>.
20. Heaven D. Why deep-learning AIs are so easy to fool. *Nature.* 2019;574(7777):163–6. <https://doi.org/10.1038/d41586-019-03013-5>.



Artificial Intelligence in Clinical Toxicology

108

Meetali Sinha, Praveen G., Deepak Kumar Sachan, and Ramakrishnan Parthasarathi

Contents

Introduction: Clinical Toxicology	1488
The Importance of Toxicovigilance	1489
Predicting Clinical Efficacy and Drug Toxicity	1490
Artificial Intelligence	1490
Machine Learning Algorithms	1491
Deep Learning Algorithms	1496
Advances in Computational Toxicology	1497
Big Data for Toxicology Interpretations	1497
Physiological-Based Pharmacokinetic (PBPK) Modeling	1498
Conclusion	1498
References	1500

Abstract

Artificial intelligence (AI), mainly machine learning and deep learning algorithms, have advanced remarkably in the multiple domains of medical sciences. Clinical toxicology is a branch of toxicology that explores the adverse effects resulting from exposure to the various

harmful chemicals on humans. In this chapter, we explored a broad understanding of AI methods with a special focus on their applications in clinical toxicology. The future of clinical development hugely relies on the convergence of the latest digital data resources as well as the advanced computing capabilities by efficiently utilizing AI and machine learning algorithms. The advancement of computational and AI-based methods for virtual screening and *in silico* drug design has made significant progress over the last decade. Also, the use of the deep neural network in conjunction with data-driven and mechanistic modeling for clinical toxicology evaluation has emerged as a promising field for research and development. We describe the fundamental concepts of clinical toxicology and AI in this

M. Sinha · D. K. Sachan · R. Parthasarathi (✉)
Computational Toxicology Facility, CSIR- Indian Institute
of Toxicology Research, Lucknow, Uttar Pradesh, India

Academy of Scientific and Innovative Research (AcSIR),
Ghaziabad, Uttar Pradesh, India
e-mail: partha.ram@iitr.res.in

P. G.
Computational Toxicology Facility, CSIR- Indian Institute
of Toxicology Research, Lucknow, Uttar Pradesh, India

chapter. Machine learning architectures are used to analyze and learn from publicly accessible biomedical and clinical trial datasets, real-world information from sensors, and health records are briefly covered. Recognition of complex patterns in toxicological data using advanced AI methods and its assistance in detection, characterization, and monitoring of clinical diseases are also elaborated. The future of AI and its impact on clinical toxicology is also discussed and summarized in detail.

Keywords

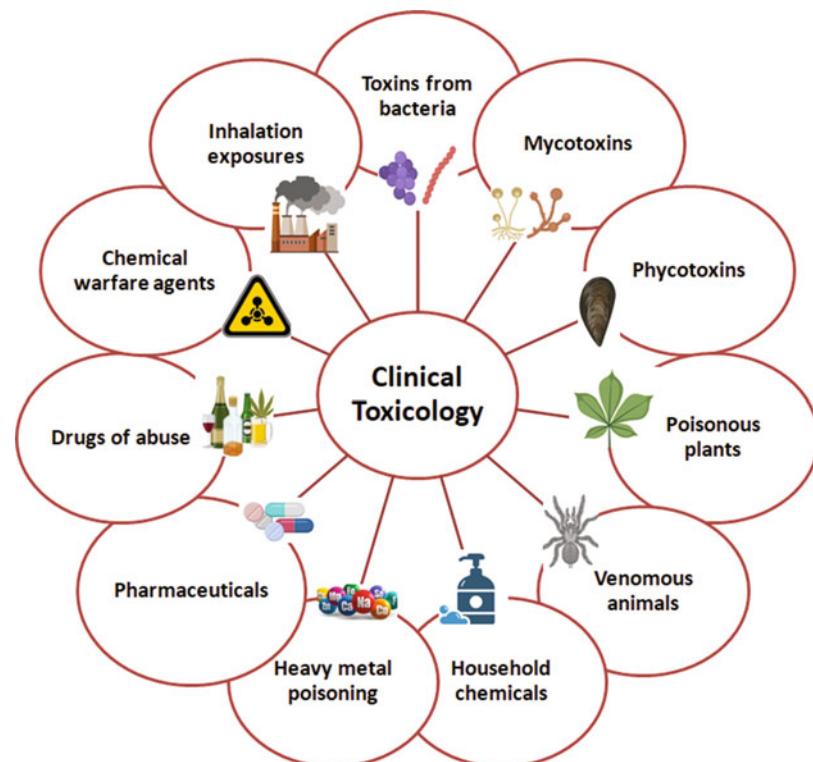
Clinical toxicology · Artificial intelligence · Deep learning · Machine learning · In silico

Introduction: Clinical Toxicology

The amplified urbanization, industrialization, and technological advancement are among the key sources of environmental health hazards [1]. Human beings are increasingly exposed to

tens of thousands of individual or mixtures of chemicals, out of which only 10% have been tested for safety [2]. It is important to understand which environmental chemical poses a risk to human health. Clinical toxicology is a sub-discipline of toxicology that emphasizes and explores the harmful effect of chemicals which are both used for therapeutic and nontherapeutic purposes [3]. It is a discipline concerned with diseases and illnesses associated with acute and chronic toxicity of chemicals. It deals with the diagnosis, treatment, and management of patients including the toxicological estimation of mildness/severity and prolonged prognosis, as well as selecting the most suitable treatment for each patient. Clinical toxicology is an amalgamation of internal medicine, pharmacology, and toxicology. Though pharmacology may deal with thousands of chemicals with therapeutic uses, clinical toxicology encompasses tens of thousands of chemicals (Fig. 1) such as bacterial toxins, phycotoxins, mycotoxins, phytotoxins, animal venoms, household toxicants, pharmaceuticals, alcohol and drug abuse [4], air/water/soil

Fig. 1 The broad umbrella of clinical toxicology, depicting the wide range of products that induce toxicity starting from the naturally occurring toxic products (produced by bacteria, fungi, algae, plants, and animals) to the synthetic chemicals (pharmaceuticals, drugs, or household chemicals). It also includes illicit and psychedelic drugs and other chemical agents that induce acute inhalation toxicity



pollutants, chemical and biological warfare agents, and many more [5, 6]. Hence clinical toxicology is branched now to its subdisciplines, including the latest toxicodynamics. This new subdiscipline is developed aiming at assisting emergency physicians with a more systematic description of the effects of the chemicals on the patients [7]. Progress in clinical toxicology is strongly interdependent on related aspects of medical disciplines and studies. Hence, the focus of a clinical toxicologist is generally on the key points that are: the importance of toxins, their corresponding toxic response, and the molecular/cellular mechanisms involved in the response. The toxic response is either imparting reversible or irreversible injury on living cells and hence itself; a detailed study to understand the mechanism of toxicity and the critical molecular interaction involved is a matter of great interest to understand the etiology. The branch of clinical toxicology has progressed immensely in the past couple of decades due to the enormous developments that happened in the field of advanced computing as well as in the advanced imaging and visualization capabilities on cellular and molecular biology. An enormous amount of chemicals could have harmful effects on human health. But not all the adverse effects of exposure to chemicals are fatal, they can be mild and therefore do not require medical attention. Diagnosis of any form of intoxication can be difficult, particularly when there is a lack of exposure history. To detect certain toxins, analytical treatments are conducted, although their diagnostic utility is often constrained. However, general medical care is often considered adequate to successfully treat asymptomatic patients [8].

The Importance of Toxicovigilance

The concept of toxicovigilance involves the early and precise detection of toxin exposure followed by the clinical validation and follow-up of toxicology and immune response in human beings. Toxicology evaluation centers are important nodal centers in this complex clinical process as the collection of primary evidence and poisoning statistics

are very critical to define the cause, incidence, and severity of poisonings occurring in the general population. The active detection of clinical adverse events related to toxic exposures can be achieved either from single case reports or by database mining. Both approaches proved effective ways to detect clinical adverse events induced by pharmaceutical products in the post-marketing phase. Toxicovigilance is essentially based on the medical assessment of acute or chronic intoxications on an individual basis, which requires validated information that epidemiologists either do not look for or cannot analyze as comprehensively on a large scale. During these years, clinical toxicologists are prioritizing their focus on early identification and risk assessment in cases of environmental and occupational toxic exposures in human beings rather than simply focusing on causative treatment and management of affected patients. Toxicovigilance is this branch of precise clinical evolution that is now accepted widely as a useful complement to other modalities of risk assessment, especially in experimental toxicology, exposure analysis, and epidemiological validation.

Clinical toxicologists include medical practitioners and scientists working in hospitals, research centers (pharmacology and toxicology departments), government agencies, or industries [9]. Clinical toxicologists have a pivotal role in the detection, prevention as well as in the treatment of chemical intoxications. The clinical expertise of a toxicologist is extremely critical in providing suggestions of treatment to patients who have been exposed to any toxic chemical, either accidentally or voluntarily. They may be consulted on any environmental, occupational, and legal aspects of any harmful chemical exposure. They are involved in collecting, assessing, and interpreting the results of the analysis performed on the poisoned patient [10]. It is also expected that clinical toxicologists will be engrossed in developing methods to manage major chemical disasters, for determining antidotes administered against chemical warfare agents [11], or even for evaluating adverse effects of pesticides and fertilizers, by determining the level and dosages as like single exposure or chronic low-level exposures. Even though some clinicians are assisted by the

conventional prediction models, yet these clinicians should have a thorough understanding of theoretical and experimental laboratory studies and must have research enthusiasm to participate in reforming their primitive clinical evaluation methods. This is an extremely daunting task, as performing clinical evaluations on environmental and chemical toxicities include:

1. Determining the patient's intrinsic vulnerability to hazardous chemicals and environmental toxins thorough evaluation and understanding of their family record, nutritional and genomic profiles, demographic peculiarities and ethnicity, etc. [12].
2. An exposure history that includes a complete overview of food and drinking water, prescribed and recreational medications, and other consumer and personal care products.
3. A brief place history that includes both the residence and workplace. It also covers the assessment of the living conditions of patients, including their exposure to traffic and other sources of air pollution.
4. Evaluation of biomarkers in different body organs to measure the long-term and short-term exposures of toxicants.
5. A detailed medical evaluation to monitor any physical symptoms and comorbidities of metabolic, neurological, reproductive, or other illnesses.

Predicting Clinical Efficacy and Drug Toxicity

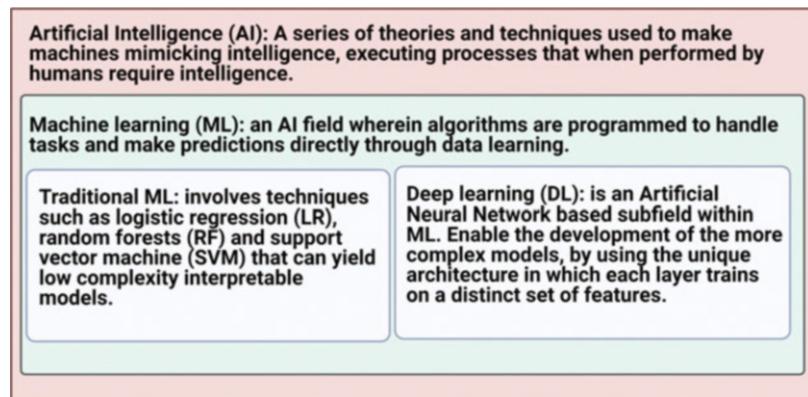
Noncompartmental analysis (NCA) is the conventional pharmacokinetic modeling tool. NCA is an experimental model that does not have a physiological interpretation component. Generally, NCA assumes whether the dose administered is distributed uniformly throughout the body and predicts the elimination of the drug through a rate constant that does not consider physiology barriers. This modeling approach definitely has some advantage like rapid generation of PK parameter estimates, but a major drawback of not considering any physiological mechanisms or biological processes that

naturally involved and influencing the actual PK dynamics. Population pharmacokinetic analysis (popPK) is yet another computational tool to determine the PK dynamics of a drug. Typically, population PK models are also experiential in design, but are different from PCA because being semi-mechanistic. There is also a PBPK model that estimates drug toxicity on the basis of physiology, biological processes, enzyme/transporter abundance, organ function, and even blood flow functions. PBPK modeling and simulation involves physiologically important parameters that are responsible for PK variability in patients.

Artificial Intelligence

Artificial intelligence (AI) is an advanced branch of computer science that deals with the realization of cognitive skills and it generally aims to mimic human intelligence with computer systems. In recent times, AI-based initiatives have attracted growing attention and acceptance in clinical evaluations like any other health care disciplines and industries [13]. The techniques of AI are widely used nowadays to improve clinical care, diagnosis, treatment, and follow-up of patients. The two subareas of AI are machine learning (ML) and deep learning (DL). ML is a subset of AI that utilizes advanced arithmetic algorithms that provide computers with the ability to automatically learn and make predictions on data without being explicitly programmed. This ability of algorithms are exploited using iterative, complex pattern matching, usually at a rate that exceeds human capability [14]. Deep learning (DL) is also a division of AI that is inspired to behave in a way similar to the human brain. It is related to understanding from examples. DL helps usually assists to filter the input data within layers and helps to predict and arrange information [15]. Recent developments in predictive DL techniques make it an ideal and useful tool on a large-scale analysis and arrangement of the newly available data; advanced storage technologies can also save significant amounts of data with advanced computing capabilities, as it can deal massive amount of data and information (Fig. 2).

Fig. 2 Distinguishing artificial intelligence (AI), machine learning (ML), and deep learning (DL). AI encompasses ML and DL wherein ML is a subset of AI and DL is a subset of ML



Deep learning appears to be more promising in terms of efficiency and precision, as it can accommodate a larger volume of data [16]. During the ML process, there is one input, one output, and one hidden layer. It requires feature extraction to train the model. On the other hand, DL uses the concept of multiple layers to adequately train the system, without the use of feature extraction. DL consists of one initial input, quite a few fully connected hidden layers, and one final output layer. Each layer represents a series of neurons and extorts higher-level features of the data, until the final layer effectively decides what the actual input shows. The additional layers the deep neural network has, the better the training of the features would be. The ability to interpret datasets differently is given by the various layers. Input data is converted by each layer into abstractions. The output layer merges those features to generate predictions. DL eventually comes into action when the required objective requires analyzing a huge number of factors correlated by unidentified and intricate interrelationships [17] (Fig. 3).

Machine Learning Algorithms

In toxicology, machine learning can be used for determining and understanding the toxicological effects of chemicals on humans [18]. These approaches provide evidence to be more beneficial as toxicity endpoints are widely heterogeneous, complex, and mostly not completely understood. Earlier several machine learning research were

mostly associated with small or nondiverse datasets that resulted in overfitting and unreliable prediction models that makes them undesirable for toxicology prediction [19]. Until recently, however, in silico approaches are considered extremely prevalent because of the advancements in computational hardware and significant concepts in techniques (like deep learning and natural language processing). The majority of these approaches have not only concentrated on using chemical structural and functional features to test chemical toxicity but have also started to implement outsized and more dissimilar sample sizes toxicity datasets that include data from both high-throughput screening (HTS) and toxicogenomics assays to supplement and explore machine learning processes appropriate for toxicology evaluation (Fig. 4).

Training and prediction are the two phases of the ML process. In the boot phase, the ML process is fed with unique datasets as input. This input dataset is separated into independent datasets:

- The training dataset: for training/learning the model. The model observes the data values/features carefully and learns from this data.
- The test dataset: for an unprejudiced evaluation of the ending model that fits on training dataset. The test set is generally a well-curated dataset that is used only after a model is fully trained. This is used for assessing the quality performance of the model (Fig. 5).

It is always beneficial to separate the datasets for the varied analysis using training

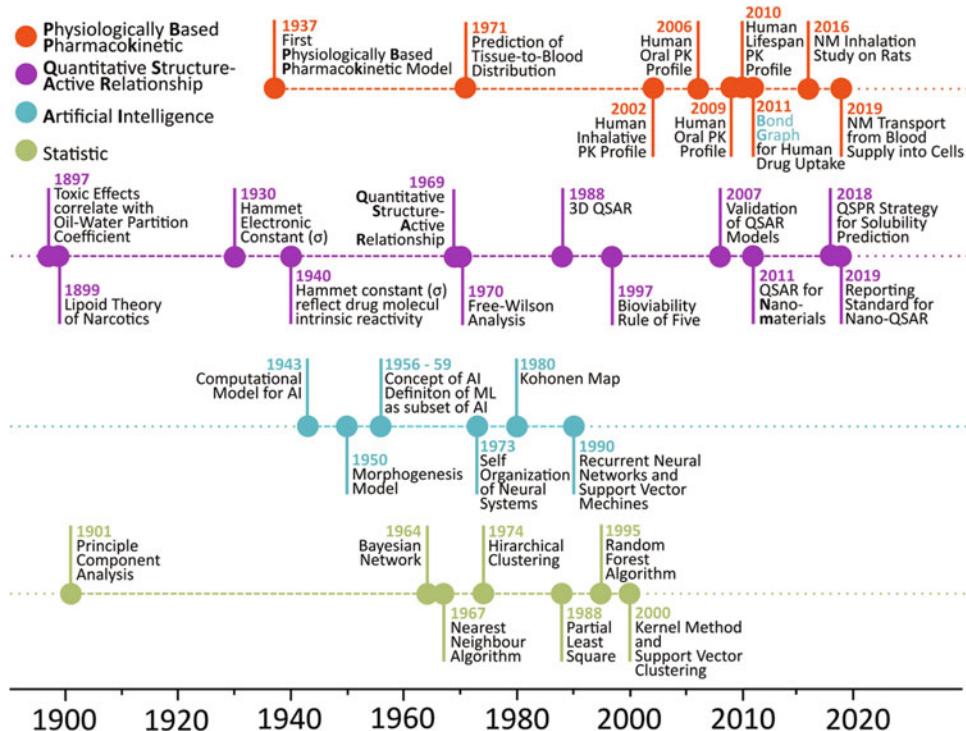


Fig. 3 Timeline of significant advances in the use of machine learning (ML) to aid in quantitative structure–activity relationships (QSAR) and Physiologically based Pharmacokinetic (PBPK) modeling and toxicity prediction of drugs and other substances. (Adapted with permission from Singh, A. V. et al., Artificial Intelligence and

Machine Learning in Computational Nanotoxicology: Unlocking and Empowering Nanomedicine. *Adv. Healthcare Mater.* 2020, 9, 1901862. Copyright © 2020 The Authors. Published by WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim)

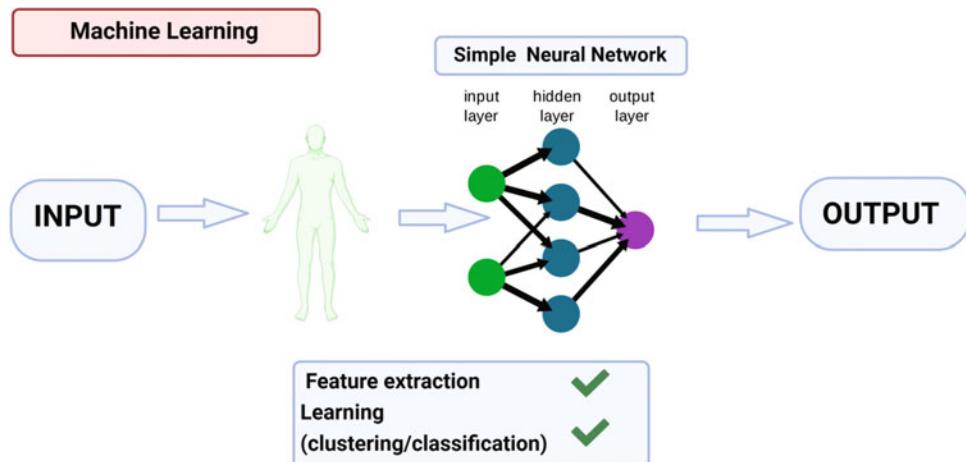


Fig. 4 Understanding the schema of machine learning algorithms

and testing models. The training dataset should not be used to validate the precision and accuracy of the model because then the model

would not generate accurate outputs if it is used in the training phase. The test dataset consists of strategically chosen data that covers

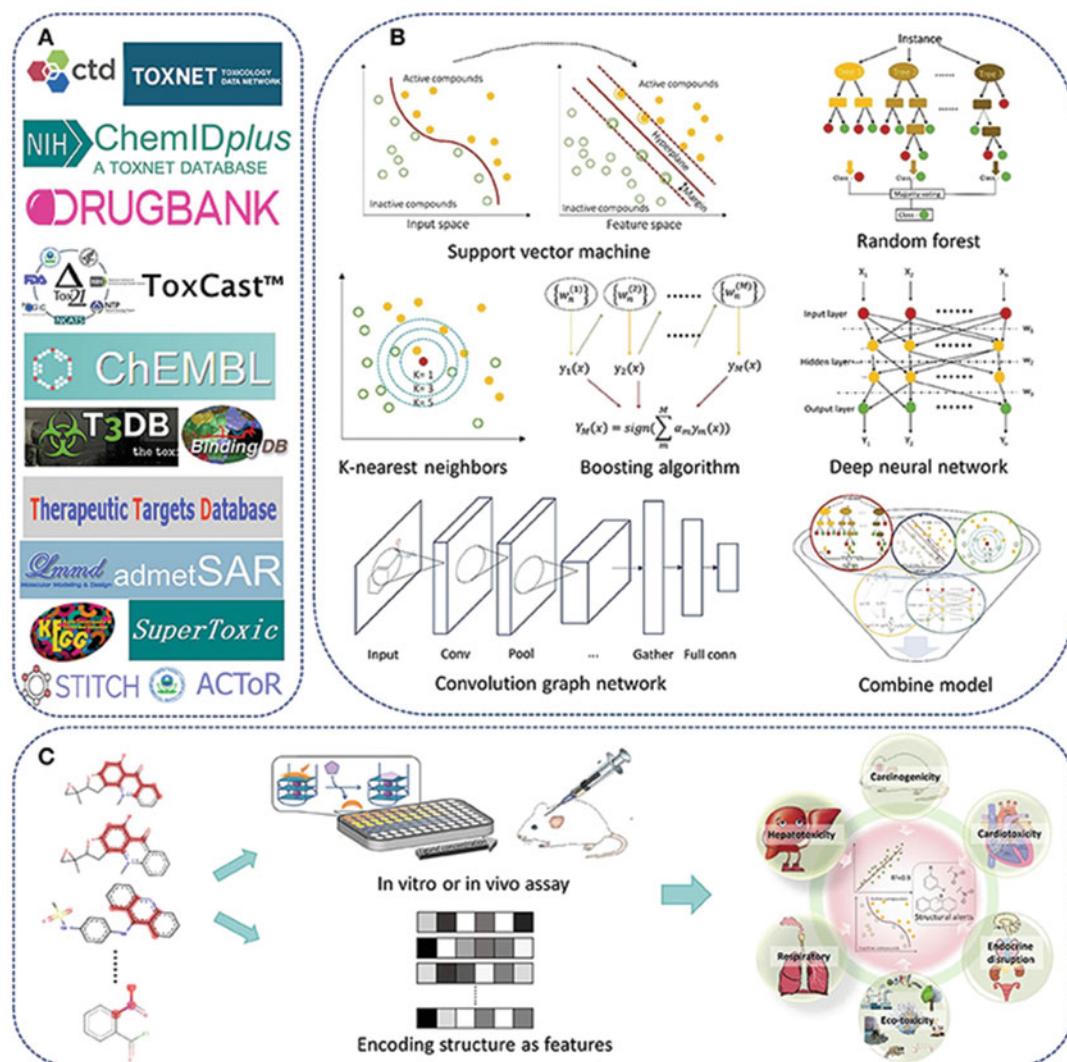


Fig. 5 In silico prediction of Chemical toxicity. The chemical toxicity prediction using machine learning techniques. A. List of the chemical databases; B. Various machine learning algorithms such as SVM, RF, KNN, etc. C. Schema of generating a QSAR model. (Adapted

with permission from Yang H et al., In Silico Prediction of Chemical Toxicity for Drug Design Using Machine Learning Methods and Structural Alerts. *Front Chem.* (2018); 6: 30. Copyright © 2018 Authors: Yang et al.)

the different class types of the original data. A typical suggested ratio to break the dataset is found to be more (nearly 70%) for the training set and almost 30% for the testing set. This splitting data method in ML is also called the “lockbox approach.” This plays a critical role in every ML project as it a determining factor of success and failure of the project. The machine learning algorithms are further divided into two forms of learning: supervised and unsupervised.

Supervised Learning

In supervised learning, predictive models are developed using a labeled training dataset (Fig. 6). To map an input to output, a supervised learning algorithm is used which learns to predict an accurate output when a new unknown input is provided. In view of toxicology, supervised learning algorithms can look over various toxicity endpoints related with chemicals as input features to classify output as toxic or nontoxic. Thus, toxicological data must be known to train the model.

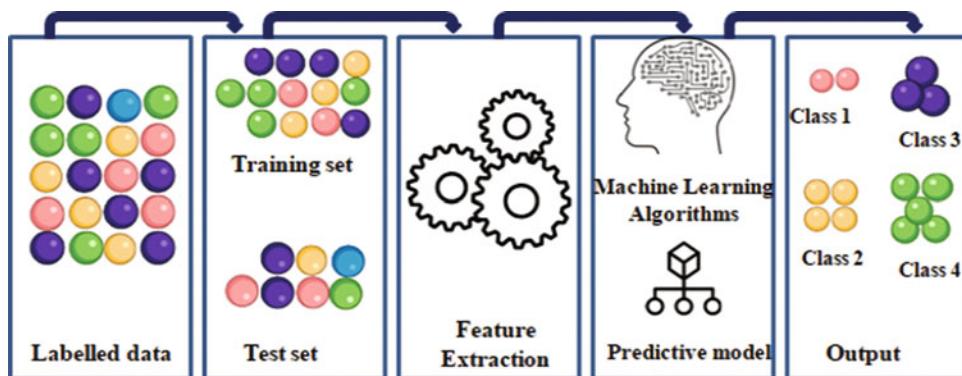


Fig. 6 Overview of the workflow of supervised machine learning

Supervised learning comprises classification and regression methods to predict qualitative or quantitative outcome, respectively. Qualitative output such as drug efficacy, organ toxicity, and disease diagnosis are streamlined using supervised classification methods, while computable output such as ADMET prediction and liver mass are modulated using supervised regression methods.

Table 1 enlists few extensively used supervised learning algorithms to test chemical toxicity. It comprises of Naïve Bayes, k-nearest neighbor (k-NN), support vector machines (SVMs), and random forest algorithms [20].

Unsupervised Learning

In the context of toxicity, unsupervised learning is useful when there is limited or almost no information available on the chemical's toxicity or non-toxicity. The purpose of the algorithm in unsupervised learning is to identify and learn representations and patterns from the unlabeled input data (Fig. 7). Unsupervised learning is very well established for the features mining, whereas supervised learning is ideal for predictive modeling by defining certain relationships within chemical characteristics as input and the result of interest as output [24].

Methods of unsupervised modeling are also used to minimize sets of feature-rich data, a feature of several sets of toxicity data, and to pick the most important predictive modeling feature or group of features. Unsupervised learning methods, for example, have been exploited to map molecular chemical bonds to generate

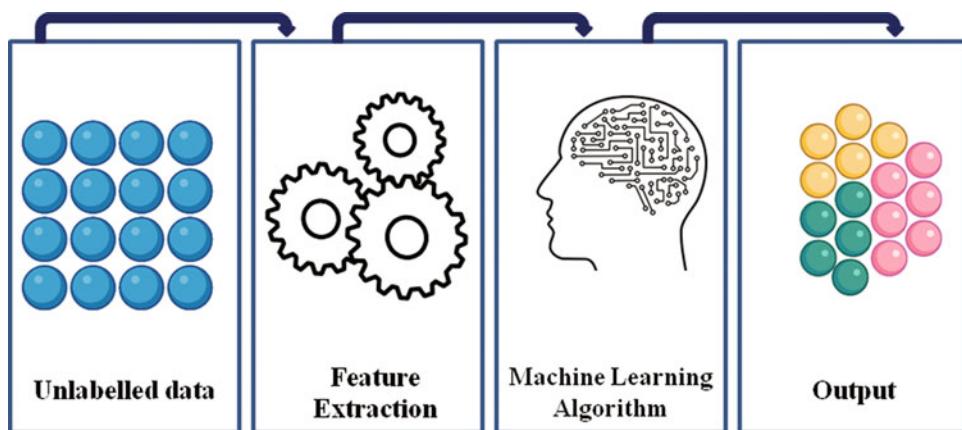
series of descriptors that are further used in combination with supervised learning to predict advanced toxicity. Clustering and dimension reduction are two major unsupervised learning methods. Table 2 enlists few of the unsupervised learning algorithms [25]. Clustering algorithms group similar characteristics into clusters together. Common clustering algorithms like k-means clustering, hierarchical clustering, and fuzzy clustering are used to cluster categorical and quantitative datasets. Another set of well-known unsupervised learning algorithms is principal component analysis (PCA) that is a well-known reduction technique used to visualize chemical patterns, reduce dimension, detect outliers, and particularly used when a trait is evaluated over a large number of dimensions, like number of genes in a genome-wide population or similar association studies.

Artificial Neural Networks (ANNs)

Artificial neural networks or connectionist systems are computing systems inspired by the biological neural networks that constitute animal brains. ANNs are a part of machine learning algorithms that mimic/replicate the learning process of neurons in the brain. ANNs consist of input, output, and a hidden layer to connect the input and output layers. However, each connection in ANN has a weight, which changes during training phase to link the input and output data. The combination and number of nonlinear connections enable ANNs to learn relatively highly complex functions to map the input and output data. This

Table 1 Description and applications of supervised machine learning algorithms

Name	Description	Applications
Naïve Bayes classifier	Based on the probabilistic Bayes theorem that is used to execute classification with a naive independence supposition between datasets It is suitable to use when the input dimensionality is high	Prediction of activity spectra for substances (PASS), an online tool, uses naïve Bayes classifier to predict various biological events, including toxicological endpoints such as carcinogenicity, mutagenicity, and developmental toxicology [21]
k-NN	A nonparametric algorithm where samples are characterized in a high-dimensional attribute by assuming that similar sets are in close proximity Selection of the feature is an imperative aspect in building k-NN models	A study by Amie D. Rodgers et al. (2010) used K-NN algorithm to predict chemical-induced hepatotoxicity. They observed that selection of total hydroxyl groups of aromatic rings had significant impact on the toxic effect of the model [22]
Support vector machines	SVM is a discriminative classifier that applies a linear or nonlinear utility to map chemical characteristics into a high degree of space and construct the best possible hyper-plane boundary which is defined by support vectors, and also helps to categorize the chemical toxicity	e-Doctor is an online application based on support vector machine that makes automatic diagnoses about health problems [23]
Random forest	It is a combined learning method where multiple resolution trees are randomly generated and combined with the majority voting schema It is more complex and less interpretable; this model performs well with large datasets and is more resistant to outlier effects and overfitting	These methods are used to categorize a broad range of toxicological endpoints such as estrogen receptor binding, cytochrome P450 inhibition, and kidney toxicity

**Fig. 7** Overview of the workflow of unsupervised machine learning

algorithm has been widely used but can suffer from overfitting [26].

The unique objective of the neural network technique is to resolve crisis in a similar process as the human brain works. During the years of advancement, importance of this technique was given to match and study specific mental abilities/illness. Divergence from biology has happened as now

capabilities are gained to acquire data on back propagation, or transition of information in the reverse direction and even to adjust the network of learning for reflecting the information. Neural networks have been used on a variety of tasks, including computer vision, speech recognition, machine translation, social network filtering, playing board and video games, and medical diagnosis.

Table 2 The learning categories and definitions for basic unsupervised learning algorithms

Algorithms	Definition	Learning category
Fuzzy clustering	A clustering method in which each data point can cluster more than once. It is a coefficient computing method that clusters for each data point	Clustering
Hierarchical clustering	A method to cluster hierarchies by combining two similar data clusters. The algorithm terminates when single cluster is left	Clustering
Principal component analysis (PCA)	A nonparametric statistical technique that utilizes an orthogonal method to convert a set of correlated features into new independent variables called principal components	Dimensionality reduction
Independent component analysis	A statistical approach for distinguishing statistically distinct additive components from a multivariable output	Dimensionality reduction
Self-organizing map (SOM)	A competitive learning network that represents the input distribution as a map by reducing the input dimensionality	Dimensionality reduction

Deep Learning Algorithms

Deep learning (also known as deep structured learning) is a type of machine learning method that uses artificial neural networks to learn representations. DL algorithms are prepared to handle very large, unstructured, and complex feature-rich toxicity data. DL supports analysis of large datasets such as high-throughput sequencing (HTS) data, toxicogenomics, microarray analysis, and image-based content [27]. Following are some of the studies conducted which have shown how deep learning models have achieved best performances in comparison to other machine learning algorithms.

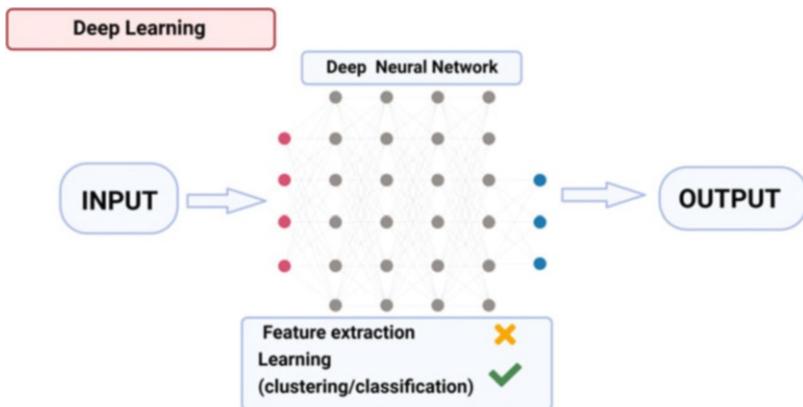
- Korotcov et al. found that deep neural network (DNN) models performed the best in a study comparing six different machine learning models among various toxicity-related endpoints for pharmaceutical research [28].
- Wang et al. assessed hepatotoxicity for biliary hyperplasia, liver necrosis, and liver fibrosis and observed that a DNN model performed better than SVM and RF models [29].
- During the Tox21 data challenge where various computational approaches for toxicity assessment was compared, a DeepTox pipeline that uses deep learning outperformed conventional machine learning methods [30].
- In a similar study that analyzed over 10,000 compounds with HTS data from Tox21, DNNs was used for predicting whether these

compounds exhibited agonistic or antagonistic activity in the androgen receptor pathway.

The majority of recent deep learning models are built on artificial neural networks, especially convolutional neural networks (CNNs), but they may also have propositional formulas or latent variables arranged layer wise in deep generative models like deep belief networks' nodes. Yoshua Bengio, Geoffrey Hinton, and Yann LeCun were most recently awarded the Turing Award in March 2019 for conceptual and engineering breakthroughs that have rendered deep neural networks a key component of computing. Hochreiter's team won the NIH's "Tox21 Data Competition" in 2014 by using deep learning to track off-target and harmful effects of environmental chemicals in foods, household goods, and medications. In 2012, a team headed by George E. Dahl won the "Merck Molecular Activity Challenge" by predicting the bimolecular target of one drug using multitask deep neural networks (Fig. 8).

An artificial neural network (ANN) with several layers between the input and output layers is known as a deep neural network (DNN). Neural networks come in a number of shapes and sizes, but they all share the same fundamental components: neurons, synapses, weights, biases, and functions. These components operate in the same way as human brains and can be conditioned much like every other machine learning algorithm.

Fig. 8 An overview of workflow in deep neural networks: DNNs are typically feed-forward networks in which data flows from the input layer to the output layer without looping back



Advances in Computational Toxicology

In the era of big data, significant advancements in computational toxicology have created an opportunity for potential toxicity research, which would have a profound effect on public health. The emergence of artificial intelligence techniques in computational toxicology is facilitating clinicians and public health professionals in bringing new research methods into practice. The latest terminologies used in the current big data era are “volume, velocity, and variety” to describe the currently available *in vitro* and *in vivo* chemical data for toxicity modeling purposes. Traditional modeling experiments consider only one entity (e.g., a single toxicity endpoint) and one set of attributes (i.e., chemical descriptors). There are several databases (e.g., PubChem, ChemIDplus, ChEMBL, DRUGBANK, T3DB, etc.) that provide a wide range of data for compounds of concern, including quantitative data directly from assays and qualitative data as the screening readout, which necessitates various data processing strategies [31].

Big Data for Toxicology Interpretations

The term “big data” refers to structured or unstructured datasets that expand exponentially and are so vast and complex that they are difficult to manage with personal computers and conventional computational methods. Big data has now become a

reality in chemical risk management. They offer multiple opportunities to transform this field of research and boost chemical risk management prospects. Big datasets necessitate specialized tools like heterogeneous and cloud computing, which go beyond traditional data collection and handling approaches, as well as complex data curation and sharing using algorithms like those used to manage data streams. The number of data is a key feature, as the term “large data” implies. The age of major toxicological evidence correlates to the evolving model of toxicology in the twenty-first century, which is turning away from apical effects research and observation and toward mechanistic toxicology. In the context of systems toxicology, the current direction of toxicology is focused on a better understanding – and continued elucidation – of the mechanisms that contribute to negative effects, and it reflects more on the individual relevance of the assessments. The number of data related to different toxicity end points is high due to the quality of the data. Combinatorial chemistry advanced quickly in the 1990s, resulting in vast chemical repositories for drug discovery screening. Automatic data mining and the use of robotics to replace humans in laboratory operations reduce the expense of testing a compound and help to quickly expand the new big data sources. Big data has thus brought major advancements to toxicology, but it has also posed significant difficulties, both on the technological side of integrating and interpreting their information, and on the general difficulty of not getting overcome by the

volume of data and being able to assess its significance. The handling, collecting, saving, combining, and reviewing of these data has become possible thanks to advances in data sciences, and data sciences has become an important part of contemporary toxicology. The distinction between experimental research and computational modeling has blurred in recent years, with the study and evaluation of certain types of data, such as omics data, necessitating the use of bespoke computational algorithms on a regular basis. In this context, predictive toxicology refers to more than just the use of statistical predictive modeling.

Physiological-Based Pharmacokinetic (PBPK) Modeling

Physiological-based pharmacokinetic (PBPK) modeling and simulation is a computer modeling technique that takes into account blood flow and organ tissue structure to determine drug pharmacokinetics (PK) [32]. The ability of PBPK models to solve ethical and technical challenges associated with pharmacokinetics and toxicology trials for various populations explains the increased reliance on PBPK models to answer regulatory and clinically relevant questions. The advantage of PBPK modeling is that it is a cost-effective and reliable statistical method that does not come with the ethical issues that come with clinical trials in vulnerable populations (e.g. cancer patients, pediatrics, pregnant women, etc.). The use of PBPK models in combination with in vitro–in vivo extrapolation (IVIVE) of absorption, distribution, metabolism, and excretion (ADME) in drug development and regulatory evaluation has grown significantly in recent years. This can be due to the creation of user-friendly platforms with high computational power that can handle this complexity and enable users to concentrate on the models' applications (Fig. 9).

Conclusion

Artificial intelligence (AI) and deep learning (DL) are gaining momentum into clinical medicine and clinical toxicology. These technologies enable to

strengthen the decision-making systems by enhancing the human intelligence. Clinicians, toxicologist, and AI developers should collaborate to take an active role in evolving their approaches and developing technology to minimize the potential adverse effects of AI in health care [33]. Integrating these programs into clinical care necessitates the formation of a mutually beneficial relationship between AI and clinicians, in which AI provides clinicians with improved efficiency and less expensive methodology, and clinicians provide AI with the clinical experiences necessary to learn complex clinical case management [34]. It will be an important task to ensure that AI does not outweigh the human face of medicine, as the public's reluctance to accept a profoundly controversial technology will be the greatest impediment to AI's widespread adoption in the future [35]. Recent advancements in AI are paving the way for proactive consultations that assess a patient's chances of acquiring a disease or other complications. Since AI can track millions of inputs at the same time, it will play a vital role in preventative health care. From a patient's digital records, AI can extract vital details. This will initially save time and increase model performance, but after proper research, it will also explicitly guide patient management. For instance, consider a consultation of a patient with an unknown complication: Currently, a physician spends a considerable amount of time reading past prescription letters, and reviewing pathological test results. Provided the patient's health record and exposure history, AI may, on the other hand, automatically plan the most critical threats and actions. It could also turn the consultation's registered conversation into a summary letter that the clinician may accept or amend. Since AI support rather than replace doctors, these applications would save a lot of time and could be introduced easily. Similarly, an AI-based mobile app capable to identify a skin lesion and diagnose its etiology, directly from the image captured, is also a future of AI in clinical toxicology. Artificial intelligence (AI) has made a lot of progress in the last decade, but it is still uncertain if these cutting-edge computing technologies will keep up with the hype. This chapter provides an

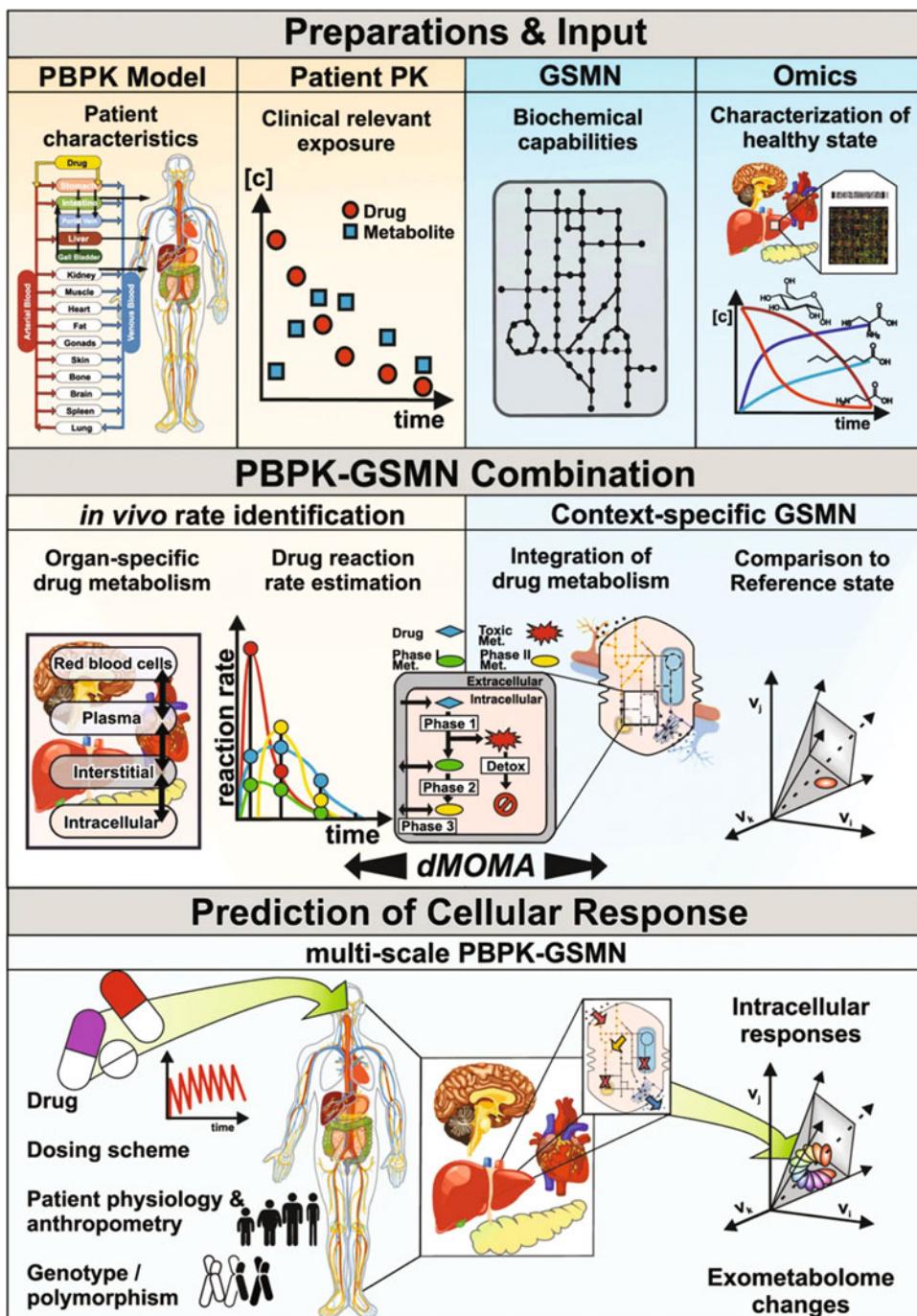


Fig. 9 Connecting PBPK models interfacing appropriate physiological organs and tissues comprising the drug-specific metabolic network describing the cellular biochemistry in the interstitial and intracellular spaces. (Adapted with permission from Singh, A. V. et al.,

Artificial Intelligence and Machine Learning in Computational Nanotoxicology: Unlocking and Empowering Nanomedicine. *Adv. Healthcare Mater.* 2020, 9, 1901862. Copyright © 2020 The Authors. Published by WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim)

interdisciplinary review of the current state of the art in using AI-based methodologies to address various clinical toxicological issues. Clinical toxicology is one field of medicine where AI systems have the ability to revolutionize the process. Researchers can use AI to predict possible toxicity or off-target outcomes in a variety of workflows, as well as process the heap of data produced from the experimental analysis quickly and efficiently in a way that is meaningful and free from human biasness. AI and machine learning (ML) algorithms can help scientists learn more about the mechanisms that cause toxicity, biomarkers, and the subtle differences in genetic makeup that affect the maximum tolerated dose or drug efficacy in different people. Finally, AI will help companies speed up preclinical growth, save time and money, and get to market faster. Apart from demonstrating superior effectiveness, emerging medical developments must also integrate with existing procedures, secure regulatory approval, and, perhaps most importantly, encourage medical personnel and patients to engage in a new paradigm.

Acknowledgments The authors thank the Council of Scientific and Industrial Research, New Delhi and CSIR-Indian Institute of Toxicology Research, Lucknow (Manuscript communication number 3749) for the research funding support and computational resources. MS thanks the Department of Science and Technology, Government of India, New Delhi for the INSPIRE fellowship.

References

- McMichael AJ. The urban environment and health in a world of increasing globalization: issues for developing countries. *Bull World Health Organ.* 2000;78: 1117–26.
- Krewski D, Acosta D Jr, Andersen M, Anderson H, Bailar JC III, Boekelheide K, et al. Toxicity testing in the 21st century: a vision and a strategy. *J Toxicol Environ Health, Part B.* 2010;13(2–4):51–138.
- Barile FA. Clinical toxicology: principles and mechanisms. CRC Press; 2010.
- Montoya ID, McCann DJ. Drugs of abuse: management of intoxication and antidotes. *Mol Clin Environ Toxicol.* 2010;100:519–41.
- Baud FJ, Houzé P. Introduction to clinical toxicology. In: An introduction to interdisciplinary toxicology. Elsevier; 2020. p. 413–28.
- Luch A. Molecular, clinical and environmental toxicology: volume 3: Environmental toxicology. Springer Science & Business Media; 2012.
- Baud F, Houzé P, Villa A, Borron S, Carli P, editors. Toxicodynnetics: a new discipline in clinical toxicology. Annales pharmaceutiques francaises. Elsevier; 2016.
- Poppenga RH. Poisous plants. *Mol Clin Environ Toxicol.* 2010;100:123–75.
- Sullivan DW, Gad S. Clinical toxicology and clinical analytical toxicology. In: Information resources in toxicology. Elsevier; 2020. p. 237–40.
- Fok H, Webb D, Sandilands E. Clinical toxicologists: the poison specialists. *BMJ.* 2016;355:i4973.
- Kuča K, Pohanka M. Chemical warfare agents. *Mol Clin Environ Toxicol.* 2010;100:543–58.
- Bijsma N, Cohen MM. Environmental chemical assessment in clinical practice: Unveiling the elephant in the room. *Int J Environ Res Public Health.* 2016;13(2):181.
- Panch T, Szolovits P, Atun R. Artificial intelligence, machine learning and health systems. *J Glob Health.* 2018;8(2):020303.
- Maddox TM. Questions for artificial intelligence in health care. *JAMA.* 2018;321:31.
- Ravì D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, et al. Deep learning for health informatics. *IEEE J Biomed Health Inform.* 2016;21(1):4–21.
- Deo RC. Machine learning in medicine. *Circulation.* 2015;132(20):1920–30.
- LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521(7553):436–44.
- Pérez Santín E, Rodríguez Solana R, González García M, García Suárez MDM, Blanco Díaz GD, Cima Cabal MD, et al. Toxicity prediction based on artificial intelligence: a multidisciplinary overview. *WIREs Comput Mol Sci.* 2021;e1516. (Early View) <https://doi.org/10.1002/wcms.1516>.
- Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med.* 2019;380(14):1347–58.
- Basile AO, Yahi A, Tatonetti NP. Artificial intelligence for drug toxicity and safety. *Trends Pharmacol Sci.* 2019;40(9):624–35.
- Parasuraman S. Prediction of activity spectra for substances. *J Pharmacol Pharmacother.* 2011;2(1):52.
- Rodgers AD, Zhu H, Fourches D, Rusyn I, Tropsha A. Modeling liver-related adverse effects of drugs using k nearest neighbor quantitative structure–activity relationship method. *Chem Res Toxicol.* 2010;23(4):724–32.
- Kampouraki A, Vassis D, Belsis P, Skourlas C. e-Doctor: A web based support vector machine for automatic medical diagnosis. *Procedia – Soc Behav Sci.* 2013;73:467–74.
- Garcia-Canadilla P, Sanchez-Martinez S, Crispí F, Bijnens B. Machine learning in fetal cardiology: what to expect. *Fetal Diagn Ther.* 2020;47(5):363–72.
- Vatansever S, Schlessinger A, Wacker D, Kaniskan HÜ, Jin J, Zhou MM, et al. Artificial intelligence and machine learning-aided drug discovery in central nervous system diseases: state-of-the-arts and future directions. *Med Res Rev.* 2020;41:1427–73.

26. Vo AH, Van Vleet TR, Gupta RR, Liguori MJ, Rao MS. An overview of machine learning and big data for drug toxicity evaluation. *Chem Res Toxicol.* 2019;33(1):20–37.
27. Chary MA, Manini AF, Boyer EW, Burns M. The role and promise of artificial intelligence in medical toxicology. *J Med Toxicol.* 2020;16:458–64.
28. Korotcov A, Tkachenko V, Russo DP, Ekins S. Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets. *Mol Pharm.* 2017;14(12):4462–75.
29. Wang H, Liu R, Schyman P, Wallqvist A. Deep neural network models for predicting chemically induced liver toxicity endpoints from transcriptomic responses. *Front Pharmacol.* 2019;10:42.
30. Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: toxicity prediction using deep learning. *Front Environ Sci.* 2016;3:80.
31. Ciallella HL, Zhu H. Advancing computational toxicology in the big data era by artificial intelligence: data-driven and mechanism-driven modeling for chemical toxicity. *Chem Res Toxicol.* 2019;32(4):536–47.
32. El-Khateeb E, Burkhill S, Murby S, Amirat H, Rostami-Hodjegan A, Ahmad A. Physiological-based pharmacokinetic modeling trends in pharmaceutical drug development over the last 20-years; in-depth analysis of applications, organizations, and platforms. *Biopharm Drug Dispos.* 2020;42:107.
33. Stead WWJ. Clinical implications and challenges of artificial intelligence and deep learning. *JAMA.* 2018;320(11):1107–8.
34. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol.* 2017;2(4):230.
35. Buch VH, Ahmed I, Maruthappu M. Artificial intelligence in medicine: current trends and future possibilities. *Br J Gen Pract.* 2018;68(668):143–4.



Artificial Intelligence in Acute Ischemic Stroke

109

Freda Werdiger, Andrew Bivard, and Mark Parsons

Contents

Introduction	1504
AI Applications to Acute Stroke Medicine	1504
Diagnosis	1505
Prediction	1510
Integration	1512
Challenges for AI in Stroke Medicine	1513
The Black Box Problem	1513
Evaluation of AI Models	1513
Data Registries	1514
Conclusion	1516
References	1516

Abstract

In recent decades, advances in image-based assessment of stroke have enabled highly effective treatments to be deployed clinically, greatly improving stroke outcomes. However, the current model of stroke care still leaves many patients without treatment for numerous reasons, including the rigid treatment time window that is often applied. Additionally, many people do not live in the range of specialist care, leaving them at greater risk of a poor outcome. Currently, artificial intelligence (AI) carries the potential to optimize stroke care by automating diagnostic processes and delivering individualized outcome predictions that can guide health care decision-making. Accordingly, there are many advances underway to implement AI into stroke care. In this chapter, a summary of

Supplementary Information: The online version of this chapter (https://doi.org/10.1007/978-3-030-64573-1_287) contains supplementary material, which is available to authorized users.

F. Werdiger (✉) · A. Bivard
Melbourne Brain Centre at Royal Melbourne Hospital,
Melbourne, VIC, Australia

Department of Medicine, University of Melbourne,
Melbourne, VIC, Australia
e-mail: Freda.Werdiger@unimelb.edu.au;
abivard@unimelb.edu.au

M. Parsons
UNSW South Western Sydney Clinical School Department
of Neurology, Liverpool Hospital, Ingham Institute for
Applied Medical Research, Liverpool, VIC, Australia
e-mail: Mark.Parsons@unsw.edu.au

AI applications to stroke medicine is presented and the challenges facing clinical deployment of AI into stroke care are discussed. Among those are data security and privacy, interpretability of algorithms, and standardization of outcome metrics. These challenges should be addressed by regulatory bodies in order to progress the field of AI in stroke.

Keywords

Artificial intelligence · Machine learning · Acute ischemic stroke · Outcome prediction · Decision assistance · Triage · Clinical outcome · Medical imaging

Introduction

Stroke is characterized by the sudden onset of neurological symptoms and is the leading cause of adult disability in the developed world. Stroke may be further classified as either ischemic or hemorrhagic. Ischemic stroke occurs when a vessel supplying blood to the brain is obstructed by a clot, or thrombus, and accounts for approximately 85% of all strokes, with a lower proportion in Asian countries. Hemorrhagic stroke occurs when a blood vessel ruptures, bleeding into the surrounding brain tissue.

Emergency care for hemorrhagic stroke focuses on management of bleeding, by keeping blood pressure down and sometimes removing parts of the skull to allow room for the brain to swell. Surgical repair may also be necessary at a later stage.

The goal of acute ischemic stroke (AIS) therapy is to either pharmacologically or mechanically remove the clot in order to restore both patency to the artery occluded by the thrombus (recanalization) and blood flow (reperfusion) to the surrounding tissue that is ischemic but still salvageable (the ischemic penumbra). The current standard of care for AIS patients is intravenous (IV) thrombolysis with recombinant tissue plasminogen activator (tPA or alteplase), restricted to less than 4.5 h from symptom onset. In addition to thrombolysis, endovascular clot retrieval (ECR)

has been strongly validated as therapy for ischemic stroke in a large vessel, typically the internal carotid artery and the middle cerebral artery, greatly improving patient outcome. ECR has one of the strongest treatment effects in modern medicine, doubling the likelihood of an individual patient living disability free after stroke [1–6].

Imaging plays an essential role in acute stroke management. Thrombolytic agents were identified as a treatment for AIS in 1958, although they were associated with a high risk of hemorrhagic complications. However, advances in neurological assessment by computed tomography (CT) scanning throughout the 1970s and 1980s led to the ability to rule out hemorrhagic stroke before the administration of thrombolytic therapies, leading to their approval by the Food and Drug Administration (FDA) in 1996 [7].

Today, most acute stroke research focuses on ischemic stroke partly due to the diagnostic dilemma present in identifying the ischemic penumbra and core, as well as the need for patient selection for ECR. With guidance from neurological imaging, clinical trials have shown that some patients benefit from reperfusion treatment up to 9 h for IV thrombolysis or 24 h for endovascular clot retrieval (ECR) [8, 9]. These advances support a more individualized form of patient assessment, rather than applying a rigid time window to all patients. By integrating artificial intelligence (AI) into image-based assessment of stroke, it may be possible to advance these diagnostic techniques to offer more optimal stroke care.

This chapter will present a summary of artificial intelligence application to stroke medicine. In section “[AI Applications to Acute Stroke Medicine](#),” the applications of AI in stroke is discussed. In section “[Challenges for AI in Stroke Medicine](#),” the challenges of AI in stroke are explored.

AI Applications to Acute Stroke Medicine

Patients presenting to a capable center with dedicated wards and specialized staff are shown to have twice the chance of an excellent clinical outcome than patients that are not admitted to stroke wards

[10]. Specialist neurologists are less likely to be recruited to remote and regional areas, putting patients in regional and remote areas at higher risk of death or severe disability from stroke [11, 12]. The most significant role of AI in stroke is to deliver the highest possible standard of care to patients, regardless of their proximity to a specialist treatment center. In seeking to do so, AI applications can be fed the same data as would be a stroke specialist, process the information and provide an output or treatment recommendation, or a recommendation to connect to a stroke specialist as a means of triage. The applications of AI for acute stroke medicine are therefore divided into three categories: *Diagnosis* (section “[Diagnosis](#)”), *Prediction* (section “[Prediction](#)”), and *Integration* (section “[Integration](#)”). Section “[Diagnosis](#)” describes methods that focus on detection of stroke and section “[Prediction](#)” addresses algorithms which serve to assist clinicians in the decision-making process by generating an individualized outcome prediction. Finally, section “[Integration](#)” discusses the integration of AI into stroke care.

Diagnosis

While histology is unparalleled as means to confirm cell death in brain tissue, MR and CT imaging is used to identify cerebral infarct in patients presenting with stroke symptoms. The gold standard for the diagnosis of a cerebral infarction is widely considered to be by an expert assisted by magnetic resonance imaging (MRI) diffusion-weighted image (DWI-MRI), which measures cytotoxic oedema [13], shown in Fig. 1. However, during an emergency MRI imaging is not always available, and computed tomography (CT) acquisitions are commonly used to diagnose by attending clinician. CT scanners are cheaper—hence more ubiquitous—and have become increasingly more compact, and mobile [14]. CT imaging modalities relevant to stroke may be summarized as follows:

- Noncontrast CT (NCCT) is a standard tool for stroke assessment. Due to increased water content, severely ischemic brain tissue appears

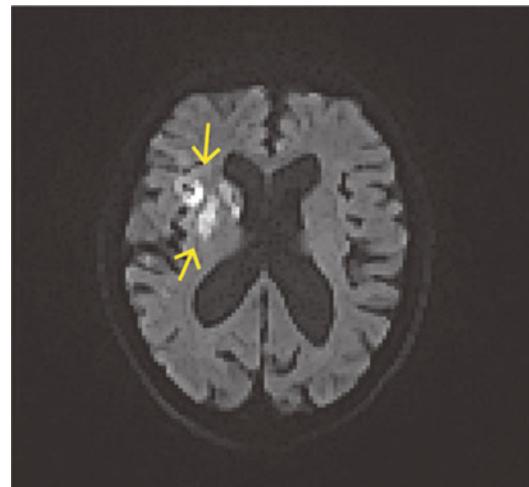


Fig. 1 MR-DWI imaging. Hyperintense regions (indicated by the yellow area) correspond to brain tissue that constitutes part of the ischemic core

hypodense on a NCCT image. This lies in contrast to intracerebral hemorrhage (ICH), which appears hyperintense on NCCT due to the high density of blood. NCCT can therefore be used to distinguish between hemorrhage and infarct.

- CT angiography (CTA) uses an injection of contrast into the blood vessels to find the precise location of the clot. This often informs treatment decisions. CTA also assists in ECR.
- CT perfusion imaging (CTP) estimates tissue response to reperfusion therapy by evaluating blood flow within the brain. Several CTA images are acquired in a short time frame and contrast concentration of tissue is measured over time. Hemodynamic maps are derived using post-processing methods that differ among software vendors [15]. Ischemic core (estimated infarction) and penumbra constitute the perfusion lesion. CTP is illustrated in Fig. 2.

Lesion Segmentation

Accurate measurement of a cerebral infarction is crucial for evaluating patients and guiding treatment. However, especially in the hyperacute phase, stroke presentation images can be subtle and often require an experienced clinician to interpret.

Manual demarcation of a stroke lesion on NCCT images is subject to inter- and intra-rater

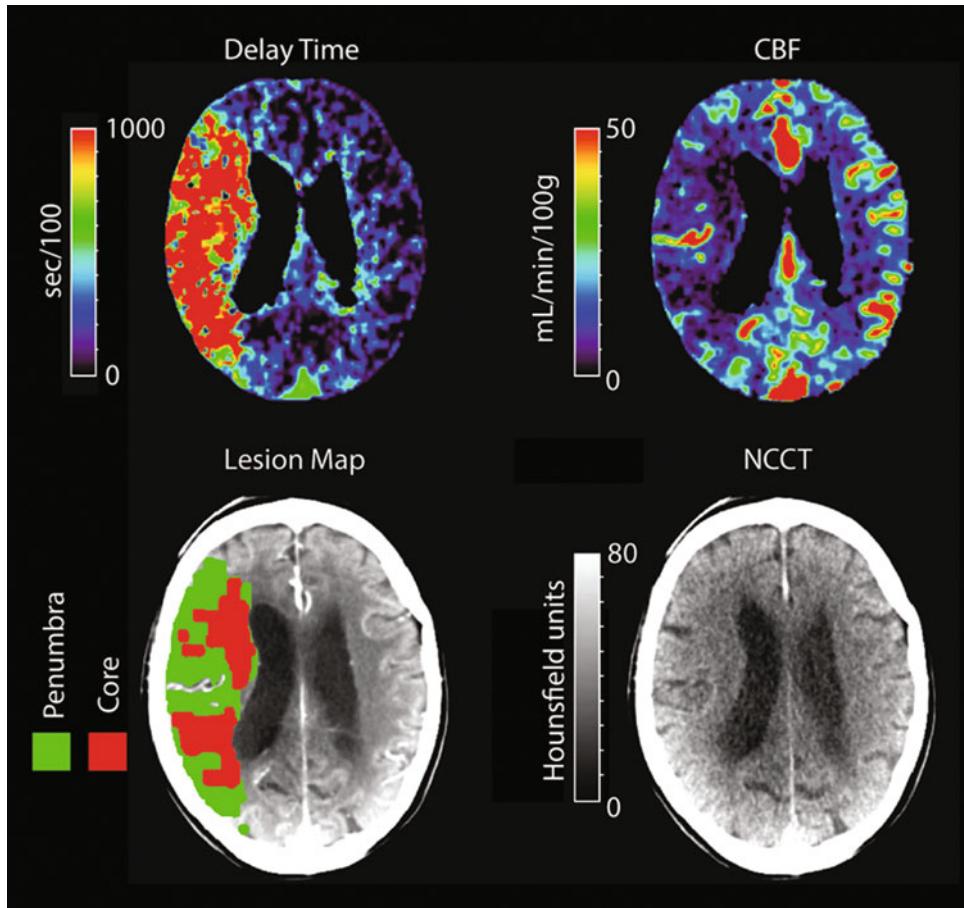


Fig. 2 CTP images. The delay time and cerebral blood flow (CBF) hemodynamic maps (generated by *Auto-MIStar*, Apollo Medical Imaging Technology, Melbourne,

Australia) provide an estimation of the ischemic core and penumbra. The corresponding noncontract CT (NCCT) is shown for the same patient

variability [16]. While a DWI hyperintense signal is visible within minutes of stroke onset, many infarcts are not visibly hypodense on a NCCT until hours after stroke onset, if at all [17]. This is partly due to the low signal-to-noise ratio which makes it more difficult to detect structural changes early on. Tissue intensities are also subject to variation due to acquisition parameters, which are not standardized. An added problem lies in the practice of windowing, done to highlight hypodense regions for easier demarcation; there is no standard placed on window levels.

CTP imaging is more sensitive to ischemic tissue. By extracting hemodynamic properties from time series data, tissue that is experiencing a deficit of blood flow can be identified. The

ischemic core identifies the region which has a degree of deficit which implies infarction and may be quantified by a 30% or more reduction in cerebral blood flow (CBF) [16]. However, other factors such as white matter disease and old infarcts have been shown to confound this thresholding approach [18, 19]. There is room for a robust method for automated lesion detection and measurement in CT-based images, validated against expertly segmented DWI.

Automated lesion segmentation is a classification problem, where each voxel (three-dimensional pixel) is classified as lesion or otherwise. Classification problems in AI may be categorized as supervised or unsupervised machine learning (ML). If annotated data (or *ground*

truth) is supplied to train a model to classify voxels into two or more known classes, the system is supervised, whereas an unsupervised model will locate patterns in the data without being supplied a ground truth. Unsupervised methods can be slow, especially over large datasets. There are also segmentation methods that combine elements from both categories, as well as those that are semi-supervised, or weakly supervised [20, 21].

The general formulation of a supervised ML classifier involves using input channels—features—to accurately predict the supplied ground truth. The objective function defined by the algorithm—which evaluates the loss accumulated by errors during a pass through the classifier—is minimized until the highest accuracy is achieved. Best practice dictates that data is divided into *training*, *testing*, and sometimes *validation* cohorts. By testing the trained model on new, unseen data, the model is prevented from overfitting to the data it was trained on.

There are several modes of approach for obtaining the necessary voxel-wise information—or features—that are used to classify the image voxel, some of which are listed here:

- *Physiological properties.* Physiological features known to be relevant—e.g., hemodynamic properties, white matter content, and NCCT hypodensity—can be quantified and calculated/extracted from images as feature maps [22, 23].
- *Template matching.* If templates are available, images can be registered to template space where deviations from template can be calculated [22, 24].
- *Statistical properties.* Information about the grey values of voxel as well as the distribution of values may be calculated [25].
- *Automated feature extraction via deep learning (DL).* Feature extraction and segmentation are combined into one end-to-end model [26].

Traditional segmentation methods characterize a voxel’s class based on the features associated with that voxel alone. In contrast, there are several methods for including spatial context into AI models—such as the location of the voxel within

the brain or in relation to other voxels—that improve the quality of the classifier.

Classifier methods such as random forest (RF), logistic regression, gradient boosting, support vector machine, or neural networks may adopt a format whereby features associated with a single instance or voxel are combined into a one-dimensional array and stacked with features from other instances to form a two-dimensional training matrix. To include spatial context with this approach, neighboring voxels can be concatenated along the feature axis. Figure 3 illustrates this patch-based approach. The voxel at the center of the patch is the target of the classification and neighboring voxels are included as additional features [24, 27].

A neighborhood analysis takes only immediately surrounding voxels into consideration and not the location of the pixel within the brain. Fully convolutional neural networks (FCNs) improve on this by forwarding the entire image through a DL network, allowing *dense inference*. Figure 4 illustrates a simplified version of the architecture described by [26]. A series of convolutional layers automatically extract features from the image. Each layer successively downsamples the image to extract features on different scales. Both fine and coarse layers are used to make predictions that respect global structure. A classification layer upsamples the image to provide a segmentation output.

The International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) has hosted the ISLES (Ischemic Stroke Lesion Segmentation) challenge since 2015, setting various tasks relating to stroke lesion segmentation. The 2018 ISLES challenge was to segment stroke lesions based on acute CTP data. The top four finalist all proposed solutions based on U-Net architecture, which adopts the FCN and modifies it to work with very few training images [28, 29]. A simplified illustration of U-Net architecture is shown in Fig. 5. The U-Net possesses an upsampling section which is more symmetric with the contracting/downsampling path, rather than the FCN’s single upsampling layer. Features from the downsampling path are concatenated with those from the upsampling path

Fig. 3 Voxel-based data with neighborhood analysis. A training matrix is constructed out of features and the ground truth class label corresponding to each voxel. The neighboring voxels are here concatenated onto the feature axis for each voxel providing spatial context

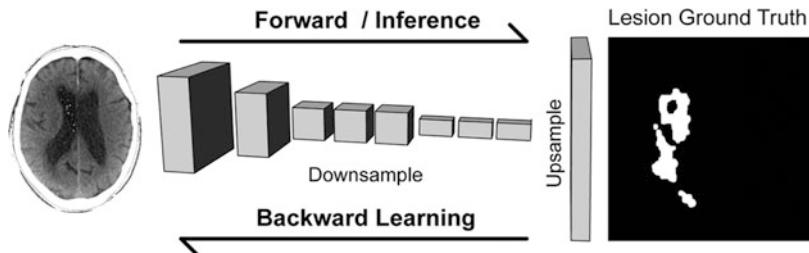
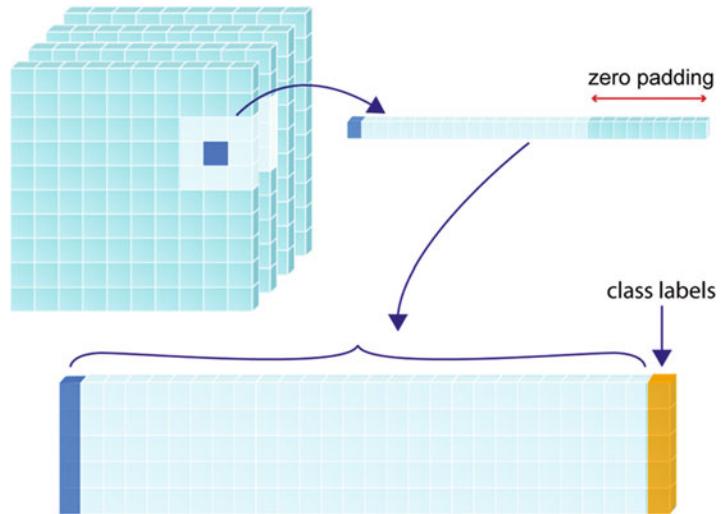


Fig. 4 A fully convolutional neural network (FCN). The entire image is forwarded through the network, allowing for *dense inference*. The process of downsampling as the images passes through convolutional layers provides a

high degree of spatial context. Features are automatically extracted over multiple scales. The image is then upsampled to provide a classification against the ground truth. Feature hierarchy is determined via back propagation

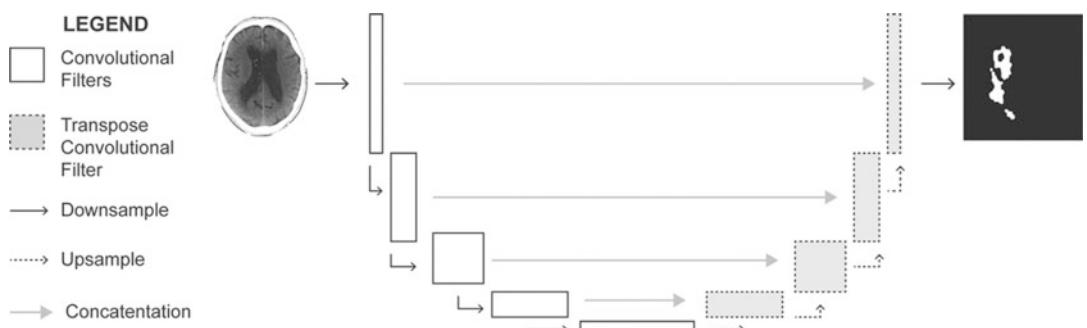


Fig. 5 The U-Net builds of the architecture of the FCN. The contracting network is supplemented by additional convolution layers, where the downsampling operators are replaced with upsampling operators, creating a symmetric U-shape. The resolution of the output is increased as

a consequence. In addition, high-resolution features from the downsampling path are added to the upsampled path, providing a more precise output. These modifications allow the U-Net to be trained by very few images

to give a high resolution and structured output. Variations on U-Net continue to be developed for the segmentation of both hemorrhagic and ischemic stroke (e.g., [30]).

Dense inference using whole images is challenging due to the computational expense. As an alternative, FCNs can be adapted to patches. Figure 6 illustrates a multiscale 3D CNN which uses two convolutional pathways, simplified from [31]. One pathway takes as its input a patch at normal resolution, and the other a larger patch at a lower resolution. Both are centered on the same image location. These two parallel convolutional pathways (each 11 layers deep in [31]) are combined and passed through two full connected layers to allow for structure prediction.

Clot Detection

Automated clot detection in CTA can be approached in several ways. Using supervised ML, a CNN may be taught to recognize a clot using annotated training data. There is FDA-approved software that reports using CNNs to detect LVOs with a sensitivity of 96%, although the details of the methods are not available due to commercial interests [32]. This approach depends

on access to a large, high-quality dataset which often has been highly selected and is not likely representative of a typical image for accuracy.

Template matching may also be of benefit to clot detection [33]. However, a CTA template relies on a high volume of high-quality images to create a generalized view of cerebral vasculature, and so this approach has a similar need for a large, high-quality dataset. It would also have difficulties due to anatomical variants that exist in brain vasculature. For example, the division of the MCA is variable after the horizontal (M1) segment, with 22% of cases branches into more than two divisions rather than a bifurcation. MCA duplication is another anatomical variant, seen in approximately 1.5% of the population [34, 35]. Nonetheless, the use of a template would allow for the automated identification of the occluded vessel by name. Stroke location together with lesion volume rather than lesion volume alone is a better predictor of functional outcome [36]. Treatment decision may be impacted if a clot lies in a region of the brain such that an infarction there would result in considerable loss of function. Small strokes in eloquent areas such as the brainstem can induce sever

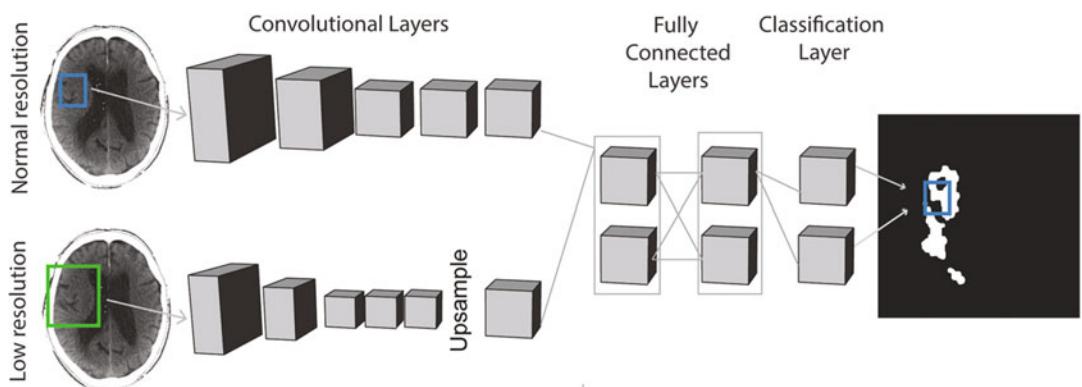


Fig. 6 3D CNN with two multiscale convolutional pathways, simplified from [31]. Using entire images as with Figs. 5 and 6 introduces issues of class imbalance and carries computational expense. This architecture adapts dense inference to patches which are randomly sampled into a training batch. To include spatial context, an additional lower-resolution patch (centered on the same image

location) is forwarded through a parallel channel. Both channels undergo a series of convolutional filters for automated feature extraction. The lower-resolution channel is upsampled to match the higher resolution. Both channels are then combined in full connected layers for structured prediction. Finally, a single classification layer provides the output

deficits comparable to those associated with a large hemispheric stroke [37].

Prediction

As stroke progresses, brain tissue that is ischemic may evolve to become infarcted and irreversibly damaged. As the volume of ischemic tissue increases, so too does the breakdown of the blood-brain barrier which increases the likelihood of hemorrhage after reperfusion. Treatment decisions may therefore be framed as a matter of risk over reward and should be made individually based on the characteristics of each patient. In Australia, only 10% of stroke patients receive IV treatment, partly due to many patients arriving at stroke centers outside the 4.5 h treatment window [38]. Therefore, there is incentive to develop AI algorithms that predict the outcome of clinical decision for individual patients so that patients can be effectively selected for treatment.

Imaging Outcomes

The volume and location of an infarction is correlated with patient outcomes and thus predicting infarct core growth can be key in clinical decision-making [39]. Penumbra imaging such as CTP provides an estimate of core growth under the assumption that the penumbra will go on to infarct if blood flow is not restored. The volumetric difference between core and penumbra, defined as the *mismatch*, is the therapeutic target and therefore drives clinical decision-making.

Although CTP imaging has been successful in selecting patients for treatment that would otherwise go untreated under the current standard of care, the therapeutic target is estimated under the assumption that treatment is provided without delay. In reality, patients may have to be transported to a treatment center before undergoing reperfusion therapy, and delays may occur within a hospital setting as well. This may lead to the growth of the infarct core, altering the core/penumbra ratio. Therefore, to have robust prediction for final infarct core, the impact of time delays should be modeled. However, infarct growth over time cannot be modeled based on a baseline

perfusion lesion alone as characteristics such as age and sex impact disease progression [22]. Additionally, fluctuations in collateral flow are known to occur, and failure of collaterals is itself associated with core growth, establishing a complex nonlinear relationship between time and infarct core growth [16, 40].

The idea of patient selection based on mismatch is also predicated on the assumption that reperfusion therapy will result in full recanalization of the affected vessels. However, treatment can also result in degrees of success [41]. Singular events such as hemorrhagic transformation may also result from treatment. There may be higher-order interactions (more than two) between characteristics that are predictive of treatment outcome which are not yet known. These complexities must be accounted for to provide robust core prediction.

It has already been established that DL can provide a more robust means of segmenting the ischemic core than with univariate thresholding. Another vital contribution of DL to stroke research is the ability to automatically examine and model complex nonlinear relationships and higher-order interactions [42, 43]. This lies in contrast to conventional statistical models, the vast majority of which rely on linear interactions. ML systems that are based on linear models, such as logistic regression, may also be too simple to predict lesion outcome [44]. The design of statistical models is also dependent on the perspective of the user, which is an obstacle if one wishes to advance beyond their understanding of a system [45]. The hidden layers of a DL network, and a well-chosen activation function (the function which determines the output of each layer) can trigger the extraction of latent (hidden) features. Complex interactions among latent features of a system can be explored to predict tissue fate.

Multiscale 3D CNNs, similar to the diagram shown in Fig. 6, can be adapted from lesion segmentation tasks to formulate lesion outcome prediction models. The dual-channel architecture described by [31] can be expanded to include additional channels for patient or clinical data that is relevant to the predicted outcome such as age, sex, time to treatment, and degree of

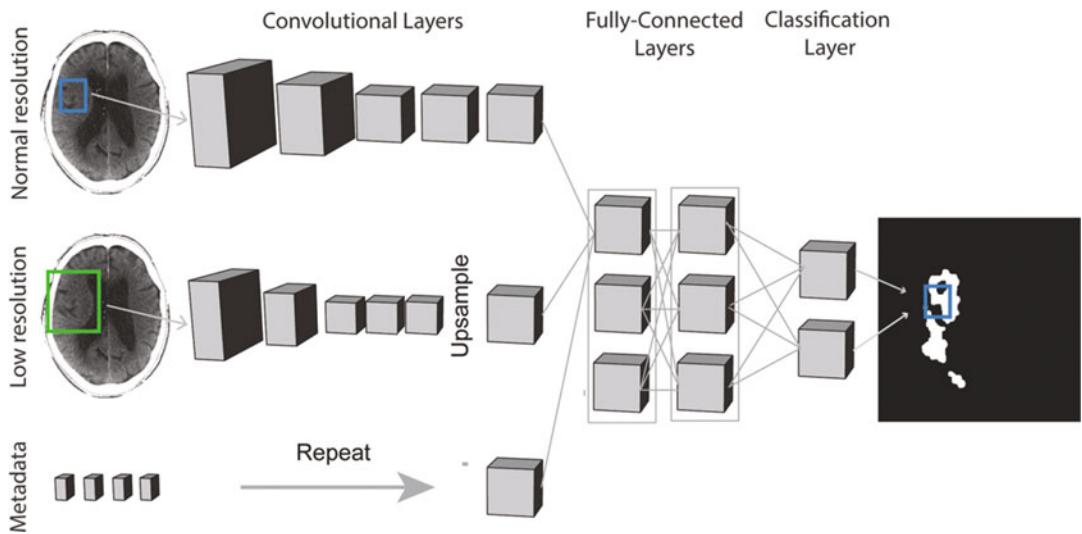


Fig. 7 Dual channel architecture (simplified in Fig. 6) may be expanded to include additional channels. Here, a third channel inputs metadata, which is repeated in an array to match the size of the normal resolution patch. All feature

channels are concatenated along the same axis and passed through fully connected layers for structured prediction. Simplified from [41], who include time and treatment success as predictors of outcome

recanalization [41]. Figure 7 shows a simplified version of the multichannel architecture described by [41].

Clinical Outcomes

While lesion properties may be associated with clinical outcomes, it is not necessarily so. The relationship between structure and function is impacted by elements beyond lesion characteristics, such as the existence of previous infarcts and white matter hyperintensities [46]. Age and stroke severity, measured by the National Institutes of Health Stroke Scale (NIHSS) score, along with other clinical variables, have significant impact of post-stroke outcomes as well [47]. Therefore, as most stroke trials use long-term functional outcomes as the primary end point, it may be suitable to directly predict a functional outcome rather than final infarct [37].

One such outcome is the modified Rankin Scale (mRS), a clinician-reported descriptor of global disability. The mRS is used often in clinical trials due to its accepted inter-rater reliability and straightforward application [48]. Briefly: mRS ranges from 0–6, where 0 is no impairment, 5 is severe disability requiring constant care; 6 is death. The mRS at 90 days after stroke is commonly used

to parameterize clinical outcome. Many AI algorithms developed to predict clinical outcome of stroke use a binarized mRS as ground truth, e.g., where $mRS < 2$ may denote a good outcome.

There are various scoring systems—based traditional statistical analysis which have been adopted by medical communities to predict the clinical outcome of acute ischemic stroke. The Totalled Health Risks in Vascular Events (THRIVE) score uses preadmission information to predict clinical outcome on a nine-point scale. It is designed to help clinicians understand a patient's chance of a good outcome. The ten-point Houston Intra-Arterial Therapy (HIAT) score uses clinical variables as well as imaging variables to patient outcome as 90 days. HIAT includes the Alberta Stroke Program Early CT score (ASPECTS), a ten-point score designed to evaluate NCCT images of patients with MCA stroke. Both THRIVE and HIAT assume a linear relationship between variables and outcomes [49]. As mentioned earlier, subjective design is a disadvantage to statistical models such as THRIVE and HIAT and that high-level ML models can investigate interactions between feature that have not considered by the user. Further, these statistical scoring methods are fixed and are not designed to adapt to added

information. ML models, on the other hand, can continue to learn and improve predictions as data examples accumulate.

ML models can adapt the nuances of patient outcome by including more features and exploring interactions between them. The 2016 ISLES challenge focused on outcome prediction, tasking challengers to predict both the lesion *and* the follow-up mRS score using 35 training cases and approximately 40 testing cases. Variables used to formulate a prediction include the time since stroke (TSS, time between stroke onset and imaging), the time to treatment (TTT, time between imaging and treatment), and degree of recanalization, parameterized by the Thrombolysis in Cerebral Infarction (TICI) score. Two out of the top three model for predicting clinical outcome adopted random forest, a high-level ML technique also capable of modeling the relevant complex interactions [44]. RFs consist of an ensemble of randomly trained decision trees, designed to increase predictive accuracy and avoid overfitting. The decision trees are trained in parallel, thus RFs train very quickly. Although most used for classification, RFs can also be formulated for nonlinear regression based on multidimensional input features [50].

There are other clinical outcomes that an algorithm can be trained to predict, depending on the application of the algorithm. Examples of clinical outcomes that are relevant to clinical decision-making include hemorrhagic transformation or successful recanalization after thrombectomy or thrombolysis [51, 52]. For patients suffering from ICH, the ability to estimate the risk of hemorrhagic expansion is valuable when selecting patients for surgical interventions. The methods that have been discussed may be capable of modeling the complex interactions between variables that lead to these outcomes and providing a prediction. There is ongoing development of techniques to predict stroke outcomes.

Integration

It has been mentioned several times already that the relevance of these algorithms will depend on the way they can be integrated into the stroke care

continuum and patient pathway. Technology has already been integrated into the stroke care workflow via the introduction of Telestroke. Rapid and reliable diagnosis can be delivered via Telestroke through remote consultation with highly trained clinicians [12]. In addition, mobile stroke units (MSUs) are being deployed in various countries throughout the world [53, 54]. These units update and equip patient transport vehicles with the tools necessary to assess for stroke, including imaging tools. Patients can therefore be provided with pre-hospital stroke care and thrombolysis.

These advances are especially relevant to areas that do not lie within proximity to stroke treatment center. Moreover, there is a growing shortfall in neurologists that is expected to rise and so there may be a growing percentage of the population that does not have access to expert stroke assessment [12]. Ageing populations will also contribute to a higher rate of stroke occurrence. Therefore, it is highly relevant to develop and optimize automated diagnostic and predictive AI algorithm so that they can be integrated into the stroke care workflow.

There is also opportunity for AI integration to assist in centers that already provide specialized stroke care. Even in stroke centers, communication remains a burden to physicians. However, with a messaging system that is integrated across all stages of patient processing—advanced emergency care, patient transportation, emergency department, and finally stroke treatment center—physicians can be alerted in advance of an incoming patient. Patient information can be collected and sent by the same messaging system. If the patient has existing electronic medical records, this additional information can be made available as well. Historical and pre-admission data, MSU imaging, and additional information such as the estimated transport time may be used by AI to assist in diagnosis and determine outcome prediction [12].

Currently, AI is already deployed clinically into Telestroke in a triage system [32]. There, all patients with suspected stroke will have their details reviewed by a neurologist. However, if the clinical and/or imaging data of a patient is first passed through an AI algorithm, their place in line can be updated based on the results of the

algorithms. For example, a positive result for a suspected LVO will place a patient ahead of others which have returned a negative result. This process has been shown, in a limited study, to reduce patient processing time.

Challenges for AI in Stroke Medicine

The Black Box Problem

A great barrier in the application of AI is its interpretability, or the degree to which the internal mechanics of a machine can be understood. The deployment of AI often invokes black box toolkits—whereby the user applies an input to retrieve an output—which is not problematic provided the algorithms in use are highly interpretable, such as decision trees or logistic regression. Neural networks, on the other hand, have a problem with interpretability. A deep neural network has a complex hidden structure such that it may be difficult to understand and interpret the internal processes and explain why a decision was made. An associated problem for AI is *explainability* which is the ability to explain to a human audience what is going on. When it comes to AI in stroke, this is troubling on two fronts.

First, there is the issue of trust. For AI processes involved in decision assistance or outcome prediction, it does not seem feasible or ethical for clinicians to rely on algorithms which have processes that are mysterious to them. This may not be a problem for AI algorithms that assist in or automated of processes that would otherwise remain as time-costly and labor-intensive work for radiologists or neurologists (e.g., denoising and segmentation). In those cases, the decision-making of clinicians is expedited. This is highly valuable in the time-pressured environment of acute stroke medicine and allows for more time with the patient and their family, to consider complex or delicate aspects of the decision-making process that are not captured by AI (e.g., cultural and financial). However, for predicting outcome and aiding in decision-making, doctors and medical boards may remain skeptical.

Secondly, there is the matter of legal responsibility if uninterpretable AI generates an incorrect

output. If one cannot understand the hidden processes of a neural network, one cannot predict how the AI will continue to develop as it learns and whether that would lead to an unexpected result. In fact, it is more or less inevitable that if AI is deployed clinically, and continues to learn with more data, it will eventually produce a decision that is inexplicable. It is important to understand how to manage that inevitability and more AI-based tools are deployed.

There are solutions to both issues of explainability and interpretability. For the former it is important that strong collaborative ties are made between developers and the eventual users of algorithms. Poorly communicated ideas result in skepticism and dubiousness. For the latter, it may be necessary for developers to add steps in the deployment of AI that render the process more transparent to both the developers and toward the users once they have been properly visualized and explained. There are many developing solutions to DL interpretability [22, 55].

Evaluation of AI Models

As it stands, there is no standard in the use of accuracy metrics to report on the performance of an AI model. Receiver-operating characteristics (ROC) are widely used to evaluate binary classification systems in medical research and are therefore often applied to evaluate the performance of such AI models in medicine. Illustrated in Fig. 8, the ROC curve plots true positive rate (TPR; also called *sensitivity*) against false positive rate (FPR; *specificity* is $1 - FPR$), and the area under the curve (ROC-AUC) often represents the overall performance of a model. The success of otherwise of the derived model in predicting a binary outcome for a new patient defines the ROC-AUC.

The interpretation of ROC performance metrics for ML models is subject to some nuances and so can be highly subjective. For instance, a binary classification, while convenient for constructing a model, may be artificial and impractical. As discussed earlier, predictive models may define a binary outcome as ground truth, such as with the binarized mRS. However, it may not be useful to know that a patient is at risk of an outcome

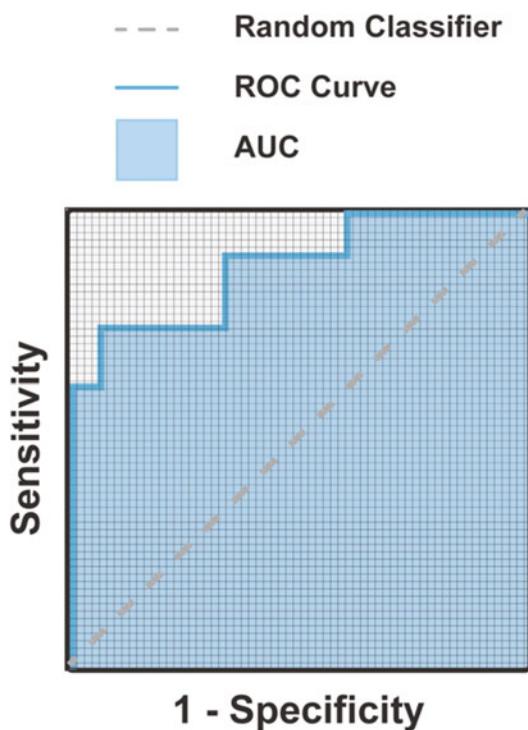


Fig. 8 Receiver operating characteristic (ROC) curve. For a binary classifier, the area under the ROC curve (AUC) determines the overall performance of the model. The broken line represents the ROC curve of a random classifier

encompassing both moderate disability (mRS = 3) and death (mRS = 6). Instead of classifying patients into simple categories to simplify a model and increase predictive values, it is more clinically relevant to precision medicine to predict a continuum of risk [56]. Similarly, diagnostic or predictive models are more applicable to clinical medicine if they can identify a multitude of classes, rather than being limited to making only one distinction [57]. Binary outcomes are not necessarily a requirement of ML models. Discriminative models can provide a multiclass rather than a binary output via *one-hot encoding* (Table 1); however, again, there is no standard in how to represent the accuracy of a model capable of predicting multiple outputs.

When interpreting the reported accuracy, the training cohort should also be considered carefully. The accuracy of a model may differ depending on characteristics of the study

population. Many AI algorithms are trained on limited or highly selective datasets, such as patients with confirmed LVOs, which constitute only 14–16% of all stroke cases. When applied to a more general population, or under different conditions, the model may not perform with the same accuracy. A model that performs with high accuracy but only under certain unrealistic conditions is of little relevance to clinical practice [56].

Related to training population is the particular matter of class imbalance within classification training data. Generally, for a classification problem, training data should have close to equal class representation so that the algorithm does not favor the majority class and learns to be accurate in its prediction of the minority class as well. However, medical image segmentation problems are typically highly unbalanced because the frequency of healthy pixel greatly outweighs that of unhealthy pixels. There are numerous solutions to this problem, such as patch sampling from each class or weighting the loss associated with the minority class [26, 31]. However, there are no clear guidelines with how to manage class imbalance.

Nevertheless, for scenarios with a large class imbalance where there are excessive number of correctly classified background voxels, the Dice similarity coefficient (DCS), Jaccard index (JI), or Hausdorff distance (HD) are popular as accuracy metrics to replace the ROC-AUC. DCS and JI are calculated based on the amount of overlap between the predicted and measured areas, while HD computes the maximal displacement between the two areas. Figure 9 illustrates the difference between DCS and JI. These reflect both size and location agreement and so are better reflective of the quality of the prediction than a pixel-wise ROC-AUC accuracy. However, these metrics are also highly dependent on the test set. Without standardized reporting parameters, it is difficult to ascertain the excellence of AI models [58].

Data Registries

There are many AI algorithms that depend on the existence of large high-quality, annotated, unbiased datasets. As mentioned in section “Evaluation of

Table 1 An illustration of one-hot encoding for four classes. Each class is translated into a binary string

Class	Codes			
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1

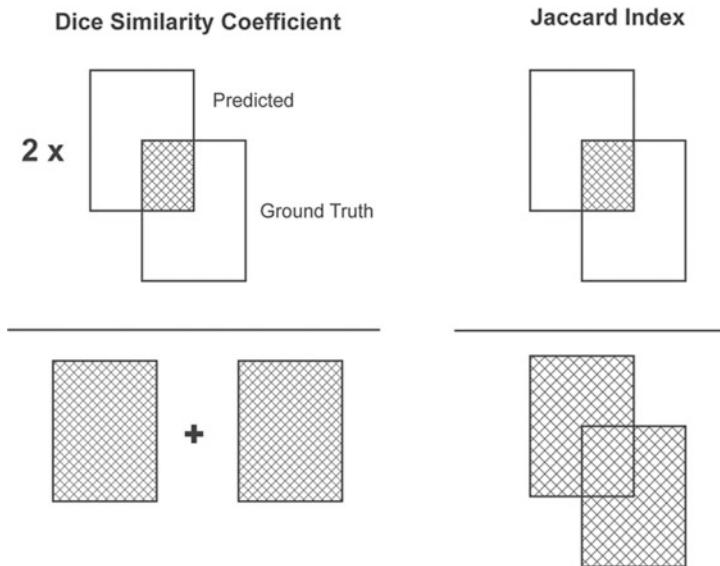


Fig. 9 The Dice similarity coefficient (DSC) and Jaccard index (JI). Both are more suitable methods to evaluate image segmentation problem as they do not weigh negative background pixels in performance evaluation. Whereas JI is Intersection over the Union (IoU) of the two areas, DSC

is twice the Intersection over the sum of the two areas. A perfect prediction gives $\text{DSC} = \text{JI} = 1$. JI is more sensitive to changes in shape. While they are positively correlated, JI penalizes incorrectly classified pixels more than DSC

AI Models,” if a dataset is too small and lacks diversity, the model may suffer from overfitting. To make a model robust, it is important to develop data sources that are as diverse as possible. However, image acquisition practices and data protocols vary among institutions, and within institutions over time due to practice changes, making it difficult to combine data sources [58]. Organizations such as ISLES supply publicly available annotated datasets for developing and evaluating AI algorithms for stroke but growing and maintaining these datasets to reflect practice changes is difficult. To do so, large numbers of images need to be scrutinized by experts, which is burdensome.

However, when discussing the need for large, readily available datasets, it is necessary to also discuss issues of data security and patient privacy.

It is seen in other areas of data-driven AI that the high quality of predictive algorithms tends to come at the cost of privacy. Because the health care industry is implied to face the highest number of data breaches among all sectors, it can be assumed that as patient data is used more frequently, privacy will become a significant issue [59]. Considerable steps will need to be taken to protect data from both external cyberattacks and internal breaches.

There are methods to adapt to limited data registries. Mentioned previously, the U-Net provides a framework that adapts to fewer training examples. Another interesting solution is the use of generative adversarial networks (GANs), which are capable of generating or augmenting data that can be used to supplement training

examples [60]. Generative models are a form of unsupervised ML, designed to automatically learn characteristics of input data such that it can produce new examples of data. By combining this with a discriminator, whose job is to classify the generated examples as “real” or “fake,” GANs can produce plausible examples of data. GANs are an exciting field which promises wide-ranging applications.

Finally, there exists a fundamental challenge in using data registries to develop predictive outcome model. As discussed throughout this chapter, ML models are structured to identify the predictors of an outcome. However, even when they are highly interpretable, ML predictive models make no assumption about *causality*, and do not enable conclusions about causality to be determined without strong assumptions [61]. To establish causality, it is necessary to estimate what would happen under another set of conditions, e.g., if a different decision was made [62]. Randomized controlled trials are designed to allow for causal inference to be established by randomizing decision-making. Both the causal pathway where the decision is made and the pathway where the decision is not made are investigated. Data registries, on the other hand, act as evidence of medical decision-making that is not randomized but based on underlying factors that may or may not be reflected in the data that is available. (For several reasons, including privacy, information may be withheld or omitted from a data registry [63]).

Discussed here, one of the strengths of high-level ML models such as RF and CNNs over traditional statistical frameworks is the ability to extract latent features not subject to user input which predict an outcome. This stands at odds with a causal model-building process, which is driven by *a priori* theory, investigating particular causal pathways that are of interest. One cannot infer causality from a model if it has not been built on an explicit causal framework. Therefore, the purpose and context of a predictive model should be specified from the outset so that it can be implemented for decision assistance.

Conclusion

The diagnostic complexities of acute ischemic stroke favor solutions that are powered by artificial intelligence. However, there are some challenges that should be overcome so that AI can be deployed effectively into clinical environment. The AI applications described here provide a service that would require extensive training when administered by a clinician. The standard to which AI is to be held will heavily impact whether these challenges are met and overcome, and it will be the role of regulatory bodies to address these challenges by enforcing standards.

References

- Berkhemer OA, Fransen PSS, Beumer D, van den Berg LA, Lingsma HF, Yoo AJ, et al. A randomized trial of Intraarterial treatment for acute ischemic stroke. *N Engl J Med.* 2015;372:11–20.
- Saver JL, Goyal M, Bonafe A, Diener H-C, Levy EI, Pereira VM, et al. Stent-retriever Thrombectomy after intravenous t-PA vs. t-PA alone in stroke. *N Engl J Med.* 2015;372:2285–95.
- Campbell BCV, Mitchell PJ, Kleinig TJ, Dewey HM, Churilov L, Yassi N, et al. Endovascular therapy for ischemic stroke with perfusion-imaging selection. *N Engl J Med.* 2015;372:1009–18.
- Jovin TG, Chamorro A, Cobo E, de Miquel MA, Molina CA, Rovira A, et al. Thrombectomy within 8 hours after symptom onset in ischemic stroke. *N Engl J Med.* 2015;372:2296–306.
- Muir KW, Ford GA, Messow C-M, Ford I, Murray A, Clifton A, et al. Endovascular therapy for acute ischaemic stroke: the pragmatic Ischaemic stroke Thrombectomy evaluation (PISTE) randomised, controlled trial. *J Neurol Neurosurg Psychiatry.* 2017;88:38–44.
- Bracard S, Ducrocq X, Mas JL, Soudant M, Oppenheim C, Moulin T, et al. Mechanical thrombectomy after intravenous alteplase versus alteplase alone after stroke (THRACE): a randomised controlled trial. *Lancet Neurol.* 2016;15:1138–47.
- Röther J, Ford GA, Thijs VNS. Thrombolytics in acute Ischaemic stroke: historical perspective and future opportunities. *Cerebrovasc Dis.* 2013;35:313–9.
- Ma H, Campbell BCV, Parsons MW, Churilov L, Levi CR, Hsu C, et al. Thrombolysis guided by perfusion imaging up to 9 hours after onset of stroke. *N Engl J Med.* 2019;380:1795–803.
- Nogueira RG, Jadhav AP, Haussen DC, Bonafe A, Budzik RF, Bhuvu P, et al. Thrombectomy 6 to 24 hours after stroke with a mismatch between deficit and infarct. *N Engl J Med.* 2018;378:11–21.

10. Tamm A, Siddiqui M, Shuaib A, Butcher K, Jassal R, Muratoglu M, et al. Impact of stroke care unit on patient outcomes in a community hospital. *Stroke.* 2014;45:211–6.
11. Prior SJ, Reeves NS, Campbell SJ. Challenges of delivering evidence-based stroke services for rural areas in Australia. *Aust J Rural Health.* 2020;28:15–21.
12. Ali F, Hamid U, Zaidat O, Bhatti D, Kalia JS. Role of artificial intelligence in TeleStroke: an overview. *Front Neurol.* 2020;11:559322.
13. Warach S, Chien D, Li W, Ronthal M, Edelman RR. Fast magnetic resonance diffusion-weighted imaging of acute human stroke. *Neurology.* 1992;42:1717.
14. Parmee RJ, Collins CM, Milne WI, Cole MT. X-ray generation using carbon nanotubes. *Nano Convergence.* 2015;2:1.
15. Mokli Y, Pfaff J, dos Santos DP, Herweh C, Nagel S. Computer-aided imaging analysis in acute ischemic stroke – background and clinical applications. *Neurol Res Pract.* 2019;1:23.
16. Goyal M, Ospel JM, Menon B, Almekhlafi M, Jayaraman M, Fiehler J, et al. Challenging the ischemic core concept in acute ischemic stroke imaging. *Stroke.* 2020;51:3147–55.
17. Wardlaw JM, Lewis SC, Dennis MS, Counsell C, McDowall M. Is visible infarction on computed tomography associated with an adverse prognosis in acute ischemic stroke? *Stroke.* 1998;29:1315–9.
18. Schirmer MD, Dalca AV, Sridharan R, Giese A-K, Donahue KL, Nardin MJ, et al. White matter hyperintensity quantification in large-scale clinical acute ischemic stroke cohorts – the MRI-GENIE study. *NeuroImage.* 2019;23:101884.
19. Rudilloso S, Laredo C, Vivancos C, Urra X, Llull L, Renú A, et al. Leukoaraiosis may confound the interpretation of CT perfusion in patients treated with mechanical Thrombectomy for acute ischemic stroke. *AJNR Am J Neuroradiol.* 2019;40:1323–9.
20. Kim W, Kanezaki A, Tanaka M. Unsupervised learning of image segmentation based on differentiable feature clustering. *IEEE Trans Image Process.* 2020;29:8055–68.
21. Pinto A, Pereira S, Meier R, Wiest R, Alves V, Reyes M, et al. Combining unsupervised and supervised learning for predicting the final stroke lesion. *Med Image Anal.* 2021;69:101888.
22. Kemmling A, Flottmann F, Forkert ND, Minnerup J, Heindel W, Thomalla G, et al. Multivariate dynamic prediction of ischemic infarction and tissue salvage as a function of time and degree of recanalization. *J Cereb Blood Flow Metab.* 2015;35:1397–405.
23. Lucas C, Kemmling A, Bouteldja N, Aulmann LF, Madany Mamlouk A, Heinrich MP. Learning to predict ischemic stroke growth on acute CT perfusion data by interpolating low-dimensional shape representations. *Front Neurol.* 2018;9:989.
24. Pustina D, Coslett HB, Turkeltaub PE, Tustison N, Schwartz MF, Avants B. Automated segmentation of chronic stroke lesions using LINDA: lesion identification with neighborhood data analysis: LINDA: auto-
- segmentation of stroke lesions. *Hum Brain Mapp.* 2016;37:1405–21.
25. Maier O, Schröder C, Forkert ND, Martinetz T, Handels H. Classifiers for ischemic stroke lesion segmentation: a comparison study. *Hu D, editor. PLoS ONE.* 2015;10:e0145118.
26. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. *IEEE Trans Pattern Anal Mach Intell.* 2017;39:640–51.
27. Zeng C, Gu L, Liu Z, Zhao S. Review of deep learning approaches for the segmentation of multiple sclerosis lesions on brain MRI. *Front Neuroinform.* 2020;14: 610967.
28. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. *Medical image computing and computer-assisted intervention – MICCAI 2015 [Internet].* Cham: Springer International Publishing; 2015. p. 234–41. [cited 2021 Mar 15]. Available from: http://link.springer.com/10.1007/978-3-319-24574-4_28.
29. Crimi A, Bakas S, Kuijf H, Keyvan F, Reyes M, van Walsum T, editors. *Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries: 4th international workshop, BrainLes 2018, Held in conjunction with MICCAI 2018, Granada, September 16, 2018, Revised selected papers, Part I [Internet].* Cham: Springer International Publishing; 2019. [cited 2021 Mar 16]. Available from: <http://link.springer.com/10.1007/978-3-030-11723-8>
30. Patel A, Schreuder FHBM, Klijn CJM, Prokop M, van Ginneken B, Marquering HA, et al. Intracerebral haemorrhage segmentation in non-contrast CT. *Sci Rep.* 2019;9:17858.
31. Kamnitsas K, Ledig C, Newcombe VFJ, Simpson JP, Kane AD, Menon DK, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal.* 2017;36:61–78.
32. Hassan AE, Ringheanu VM, Rabah RR, Preston L, Tekle WG, Qureshi AI. Early experience utilizing artificial intelligence shows significant reduction in transfer times and length of stay in a hub and spoke model. *Interv Neuroradiol.* 2020;26:615–22.
33. Amukotuwa SA, Straka M, Dehkharghani S, Bammer R. Fast automatic detection of large vessel occlusions on CT angiography. *Stroke.* 2019;50:3431–8.
34. Teal JS, Rumbaugh CL, Bergeron RT, Segall HD. Anomalies of the middle cerebral artery: accessory artery, duplication, and early bifurcation. *Am J Roentgenol.* 1973;118:567–75.
35. Komiyama M, Nakajima H, Nishikawa M, Yasui T. Middle cerebral artery variations: duplicated and accessory arteries. *AJNR Am J Neuroradiol.* 1998;19:45–9.
36. Menezes NM, Ay H, Wang Zhu M, Lopez CJ, Singhal AB, Karonen JO, et al. The real estate factor: quantifying the impact of infarct location on stroke severity. *Stroke.* 2007;38:194–7.
37. Etherton MR, Rost NS, Wu O. Infarct topography and functional outcomes. *J Cereb Blood Flow Metab.* 2018;38:1517–32.

38. National Stroke Audit – Acute Services Report 2019. Melbourne: Stroke Foundation.
39. Boers AMM, Jansen IGH, Beenen LFM, Devlin TG, San Roman L, Heo JH, et al. Association of follow-up infarct volume with functional outcome in acute ischemic stroke: a pooled analysis of seven randomized trials. *J NeuroIntervent Surg.* 2018;10:1137–42.
40. Campbell BC, Christensen S, Tress BM, Churilov L, Desmond PM, Parsons MW, et al. Failure of collateral blood flow is associated with infarct growth in ischemic stroke. *J Cereb Blood Flow Metab.* 2013;33: 1168–72.
41. Robben D, Boers AMM, Marquering HA, Langezaal LLCM, Roos YBDEM, van Oostenbrugge RJ, et al. Prediction of final infarct volume from native CT perfusion and treatment parameters using deep learning. *Med Image Anal.* 2020;59:101589.
42. Somers MJ, Casal JC. Using artificial neural networks to model nonlinearity: the case of the job satisfaction – job performance relationship. *Organ Res Methods.* 2009;12:403–17.
43. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE.* 1998;86:2278–324.
44. Crimi A, Menze B, Maier O, Reyes M, Winzeck S, Handels H, editors. Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries: second international workshop, BrainLes 2016, with the challenges on BRATS, ISLES and mTOP 2016, Held in Conjunction with MICCAI 2016, Athens, October 17, 2016, Revised selected papers [Internet]. Cham: Springer International Publishing; 2016. [cited 2021 Mar 9]. Available from: <http://link.springer.com/10.1007/978-3-319-55524-9>
45. Ryo M, Rillig MC. Statistically reinforced machine learning for nonlinear patterns and variable interactions. *Ecosphere.* 2017;8:e01976.
46. Mitra J, Bourgeat P, Fripp J, Ghose S, Rose S, Salvado O, et al. Lesion segmentation from multimodal MRI using random forest following ischemic stroke. *NeuroImage.* 2014;98:324–35.
47. Weimar C, König IR, Kraywinkel K, Ziegler A, Diener HC. Age and National Institutes of Health stroke scale score within 6 hours after onset are accurate predictors of outcome after cerebral ischemia: development and external validation of prognostic models. *Stroke.* 2004;35:158–62.
48. Banks JL, Marotta CA. Outcomes validity and reliability of the modified Rankin scale: implications for stroke clinical trials: a literature review and synthesis. *Stroke.* 2007;38:1091–6.
49. Li X, Pan X, Jiang C, Wu M, Liu Y, Wang F, et al. Predicting 6-month unfavorable outcome of acute ischemic stroke using machine learning. *Front Neurol.* 2020;11:539509.
50. Criminisi A, Shotton J. Regression forests. In: Criminisi A, Shotton J, editors. Decision forests for computer vision and medical image analysis [Internet]. London: Springer; 2013. p. 47–58. https://doi.org/10.1007/978-1-4471-4929-3_5.
51. Hofmeister J, Bernava G, Rosi A, Vargas MI, Carrera E, Montet X, et al. Clot-based Radiomics predict a mechanical Thrombectomy strategy for successful recanalization in acute ischemic stroke. *Stroke.* 2020;51:2488–94.
52. Chung C-C, Chan L, Bamodu OA, Hong C-T, Chiu H-W. Artificial neural network based prediction of postthrombolysis intracerebral hemorrhage and death. *Sci Rep.* 2020;10:20501.
53. Zhao H, Coote S, Easton D, Langenberg F, Stephenson M, Smith K, et al. Melbourne Mobile stroke unit and reperfusion therapy: greater clinical impact of Thrombectomy than thrombolysis. *Stroke.* 2020;51:922–30.
54. Bache KG, Grotta JC. Improving stroke treatment and outcomes with Mobile stroke units. *JAMA.* 2021;325:441.
55. Lee H, Yune S, Mansouri M, Kim M, Tajmir SH, Guerrier CE, et al. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nat Biomed Eng.* 2019;3:173–82.
56. Wilkinson J, Arnold KF, Murray EJ, van Smeden M, Carr K, Sippy R, et al. Time to reality check the promises of machine learning-powered precision medicine. *Lancet Digital Health.* 2020;2:e677–80.
57. Monteiro M, Newcombe VFJ, Mathieu F, Adatia K, Kamnitsas K, Ferrante E, et al. Multiclass semantic segmentation and quantification of traumatic brain injury lesions on head CT using deep learning: an algorithm development and multicentre validation study. *Lancet Digital Health.* 2020;2:e314–22.
58. Gupta R, Krishnam SP, Schaefer PW, Lev MH, Gonzalez RG. An East Coast perspective on artificial intelligence and machine learning. *Neuroimaging Clin N Am.* 2020;30:467–78.
59. Pesapane F, Volonté C, Codari M, Sardanelli F. Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. *Insights Imaging.* 2018;9:745–53.
60. Qasim AB, Ezhov I, Shit S, Schoppe O, Paetzold JC, Sekuboyina A, et al. Red-GAN: attacking class imbalance via conditioned generation. Yet another medical imaging perspective. arXiv:200410734 [cs, eess] [Internet]. 2020. [cited 2021 Mar 9]; Available from: <http://arxiv.org/abs/2004.10734>
61. Arnold KF, Davies V, de Kamps M, Tennant PWG, Mbotwa J, Gilthorpe MS. Reflection on modern methods: generalized linear models for prognosis and intervention—theory, practice and implications for machine learning. *Int J Epidemiol.* 2021;49:2074–82.
62. Hernán MA, Hsu J, Healy B. A second chance to get causal inference right: a classification of data science tasks. *Chance.* 2019;32:42–9.
63. Char DS, Shah NH, Magnus D. Implementing machine learning in health care – addressing ethical challenges. *N Engl J Med.* 2018;378:981–3.



Artificial Intelligence and Deep Learning in Ophthalmology

110

Zhaoran Wang, Pearse A. Keane, Michael Chiang,
Carol Y. Cheung, Tien Yin Wong, and Daniel Shu Wei Ting

Contents

Introduction	1520
Essential Concepts and Components in an AI System	1521
Application of AI and DL Algorithms in Ophthalmology Using Different Devices	1521

Z. Wang

Duke-NUS Medical School, National University of Singapore, Singapore, Singapore

P. A. Keane

NIHR Biomedical Research Centre for Ophthalmology,
Moorfields Eye Hospital NHS Foundation Trust, London, UK

M. Chiang

Departments of Ophthalmology and Medical Informatics
and Clinical Epidemiology, Casey Eye Institute, Oregon
Health and Science University, Portland, OR, USA

C. Y. Cheung

Department of Ophthalmology and Visual Sciences, The
Chinese University of Hong Kong, Hong Kong, China

T. Y. Wong

Duke-NUS Medical School, National University of
Singapore, Singapore, Singapore

Singapore Eye Research Institute, Singapore National Eye
Center, Singapore, Singapore

e-mail: tien_yin_wong@duke-nus.edu.sg

D. S. W. Ting (✉)

Duke-NUS Medical School, National University of
Singapore, Singapore, Singapore

Moorfields Eye Hospital, London, UK

Singapore Eye Research Institute, Singapore National Eye
Center, Singapore, Singapore

e-mail: daniel.ting.s.w@singhealth.com.sg

© Springer Nature Switzerland AG 2022

N. Lidströmer, H. Ashrafiyan (eds.), *Artificial Intelligence in Medicine*,
https://doi.org/10.1007/978-3-030-64573-1_200

1519

Retinal Fundus Photographs	1521
Optical Coherence Tomography	1535
Visual Fields	1539
Infantile Facial Video Recording	1540
Electronic Health Records	1540
Image Quality Assessment	1540
Future Research and Challenges	1542
Novel Technical Approaches	1542
Research Ethics and Artificial Images	1543
Data Ownership and Sharing	1543
Patients and Physicians Acceptance	1544
Education	1544
Guidelines	1544
Conclusion	1545
References	1546

Abstract

Artificial intelligence (AI), in particular deep learning (DL), has gained significant interest recently from healthcare systems. DL has been widely applied to detect and classify major diseases in ophthalmology, including diabetic retinopathy (DR), age-related macular degeneration (AMD), glaucoma, and retinopathy of prematurity based on fundus photographs; cataract and anterior segment diseases, glaucoma, and retinal diseases based on optical coherence tomography (OCT) scans; and glaucoma progression based on visual fields. The substantial progress of AI in ophthalmology has involved the identification of clear public health (e.g., DR screening) and clinical (e.g., prediction of the need to treat AMD) unmet needs, the targeted development of the AI algorithms using both retrospective and prospective clinical and imaging data, and designing the application interface for clinical deployment. In ophthalmology, there has also been significant experience of applying AI algorithms in “real-world” clinical situations, as well as the submission and approval by governmental regulatory agencies (e.g., Food and Drug Administration). Future research is warranted to address not only technical issues (e.g., explainability of the “black box”) but also a range of nontechnical challenges, such as increasing the awareness and acceptance of physician and patient, issues relating to global collaboration and data sharing, medical ethics, financial and reimbursement systems, and

integration of AI algorithms in clinical settings with diverse electronic health records.

Keywords

Artificial intelligence · Machine learning · Deep learning · Diabetic retinopathy · Glaucoma · Retinopathy of prematurity · Age-related macular degeneration · Electronic health records · Fundus imaging · Optical coherence tomography

Introduction

Artificial intelligence (AI) has gained remarkable traction in different segments of society, including healthcare, in recent years. This is made possible with the convergence of several interrelated technologies, such as availability of graphic processing units (GPUs), advances in computer science and mathematical models, and the increasing access to big datasets [1]. In particular, deep learning (DL), a subtype of AI, utilizing multiple processing layers to learn representation of data with multiple levels of abstraction, has shown substantial promise [2]. In some areas, DL has shown performance significantly superior to humans [3, 4] and thus DL is now widely adopted in image recognition, speech recognition, and natural language processing [1].

In ophthalmology, there have been three major areas of progress in which AI and DL algorithms

have been developed successfully. First, DL systems have been shown to accurately detect diabetic retinopathy (DR), [3–6] glaucoma, [5, 7] age-related macular degeneration (AMD), [5, 8, 9] retinopathy of prematurity (ROP), [10] predict refractive error, [11] and papilledema [12] from digital fundus photographs. Systemic diseases such as risk factors of cardiovascular disease have also been accurately predicted from fundus photographs [13, 14]. Second, there is major progress in applying DL to another imaging modality, the optical coherence tomography (OCT) scans, to detect subgroups of retinal conditions such as choroidal neovascularization (CNV) in neovascular (“wet”) AMD and diabetic macular edema (DME) [2, 15, 16]. Third, there are cases of DL algorithms used in “real world,” such as in DR screening programs and approval by regulatory agencies (e.g., FDA) in using AI for clinical care. Nevertheless, significant challenges remain. Future research is important not only to address the technical issues (e.g., explainability of the AI “black box”), but also a range of nontechnical challenges in increasing physicians’ and patients’ awareness and acceptance, and issues relating to global collaboration and data sharing, medical ethics, financial and reimbursement systems, and integration of AI algorithms in clinical settings with diverse electronic health records. This chapter provides an overview of the developments of AI in ophthalmology, highlighting major progress and future directions, including global guidelines.

Essential Concepts and Components in an AI System

AI research can be broadly divided into the clinical and technical components (Table 1). First, it is important to understand the specific clinical or research question, and then apply the appropriate AI technical methodologies to address the question [17]. For example, most AI systems in medicine have been designed to address a public health, clinical, or healthcare systems gap. AI could improve access to screening (i.e., addressing public health need), improve the diagnostic accuracy of detecting a disease or predicting the outcome of a new treatment (i.e.,

clinical needs), or relieve the shortages of specific healthcare manpower (i.e., healthcare system needs).

There are a number of factors affecting how an AI system is built. Clinical datasets and technical architectures are the two core components required to build an AI system. For clinical datasets, it is important to pay attention to the image-related factors such as types of devices (fundus photographs or OCTs), image quality, number of fields, width of field, color vs monochromaticity, classifications, reference standard, and number and experience of graders. For technical architecture, the operational flow diagram, types of input data, convolutional neural network (CNN), pre-image processing steps, and classification output and visualization map (Figs. 1 and 2) are important components.

To ascertain the diagnostic performance of an AI system, the operating threshold needs to be fixed using the training datasets. For the testing datasets, it is important to perform the power calculation by taking into account alpha and beta errors, confidence intervals, and prevalence rate to ensure that the sample size required is sufficient to confidently differentiate positive versus negative cases. The operating threshold is determined based on the universally accepted standard, tailored to the local population need. Using the same operating threshold determined from the training datasets, the area under the receivers’ operating curve (AUC), sensitivity, and specificity (with 95% confidence interval) can be determined.

Application of AI and DL Algorithms in Ophthalmology Using Different Devices

Retinal Fundus Photographs

Diabetic Retinopathy

DR is a major public health challenge [18, 19]. One-third of people suffering from diabetes is projected to develop DR, resulting in an estimated 250 million patients with DR by 2035 [20]. The early detection of DR associated with timely treatment has been largely shown to have

Table 1 The important components of an artificial intelligence system for ophthalmology [143]

Introduction	Validity of the research question Comprehensive literature search of similar technologies related to the specific disease Clinical unmet need “Value-add” of the proposed AI system
Methods	
Core components	Clinical datasets and technical network
Clinical datasets	Division of training, validation, and testing datasets
Dataset descriptions	1. Number of images, eyes, and patients 2. Inclusion and exclusion criteria for these patients 3. Study design (prospective vs retrospective) and patients demographics (optional) 4. Recruitment methods (consecutive and randomized) and sites 5. Prevalence of positive vs control cases 6. Types input data – clinical data, imaging test, or others
Technical methodology	1. Technical approach (deep learning, machine learning, or statistical approach) 2. Types of neural network 3. Operational flow of an AI system
Assessment of the diagnostic performance	1. Power calculation of the testing datasets 2. Receivers' operating curve (AUC), sensitivity, and specificity (with 95% confidence interval) 3. Accuracy, positive predictive, or negative predictive value 4. Cohen's kappa
Reference standard	Numbers and experience of graders (e.g., graders from reading centers, retinal specialists, etc.) Disease classification system
Statistical analysis and results (all with 95% CI)	1. Area under receiver's operating curve (AUC) 2. Sensitivity and specificity 3. Accuracy, positive predictive value, and negative predictive value 4. Cohen Kappa 5. Mean absolute error (continuous variables) 6. Dice coefficients (for segmentation tasks)
Discussion	1. Clinical application of the AI solution
Clinical translational value	2. Limitation of the AI systems 3. Potential deployment methods

the potential to prevent vision loss through population screening [18]. Nonetheless, the implementation of DR screening programs faces challenges related to long-term financial sustainability, as well as availability of experienced human assessors [18]. To tackle this growing crisis, a DR screening tool capable of dealing with high workload in a quick manner is critical to overcoming the limitations of DR screening programs [21]. Many computer-aided algorithms for automated detection of DR from retinal fundus images have been explored, representing potential cost-

effective alternatives to DR screening if coupled with telemedicine [22]. However, diagnostic performance comparable to humans have been reached only with DL technology [1].

Many studies using DL technology have shown robust diagnostic performance in detecting referable DR, defined as worse than mild non-proliferative DR (Table 2). For DL systems, Abramoff et al., [23] Gulshan et al., [4] Ting et al., [5] and others [6, 24, 25] have reported AUC of >0.90, and sensitivity and specificity of >90% in detecting referable DR in multiethnic

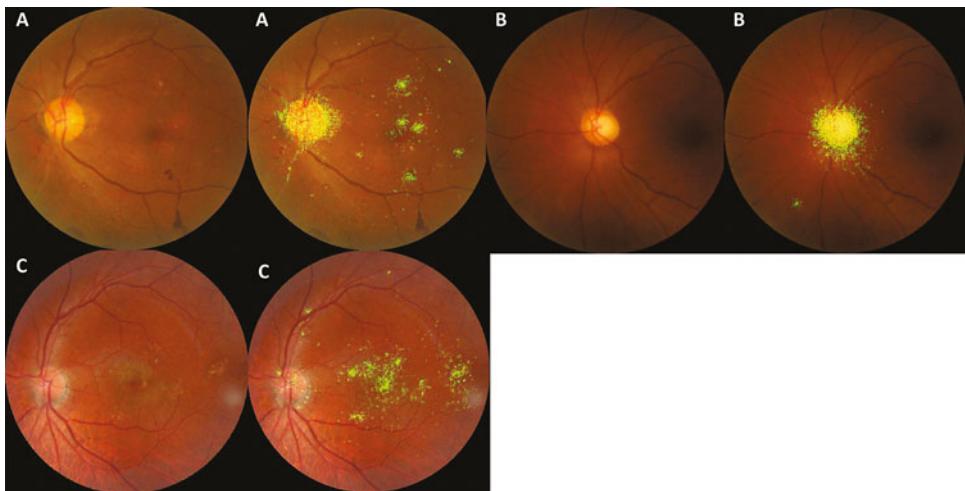


Fig. 1 Visualization techniques on fundus photos: A) proliferative DR; B) glaucoma; and C) advanced age-related macular degeneration

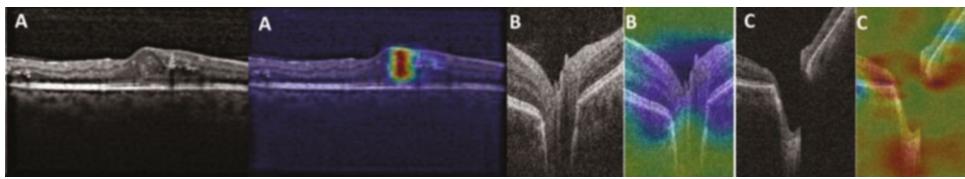


Fig. 2 Visualization techniques on OCT: A) fovea-involving DME with cystoid macular edema and hard exudates on OCT; B) normal OCT nerve fiber layer; and C) OCT nerve fiber layer with advanced glaucoma

populations with diabetes, using publicly available retrospective datasets and numerous clinic-based and population-based datasets. Following this, Abramoff et al. conducted a preregistered prospective clinical trial on a DL system, achieving 87% sensitivity and 90% specificity in detecting referable DR (worse than any DR) and subsequently obtaining US FDA approval [26]. In Australia, Keel et al. [27] also showed excellent result on a prospective AI study for DR screening in two urban endocrinology outpatient services, not only with a good diagnostic performance – 92% sensitivity and 94% specificity in detecting referable DR, but also well received by the patients (96% of participants were satisfied with the AI algorithm). Real-world performance of a DL system on detection of referable DR in prospective datasets was examined recently by Bhuiyan et al. Using nonmydriatic retinal images collected from 974 patients with diabetes in

primary setting, the DL screening system achieved a sensitivity of 82.6% and a specificity of 93.7%, which meet the requirement of FDA approval of a screening program. This DL system was also integrated with a telemedicine platform, capable of generating a report on the referral decision automatically within one minute [28]. Besides detection of referable DR, DL algorithm has been utilized to predict OCT-derived DME from two-dimensional (2D) fundus photographs with an AUC of 0.89, sensitivity of 85%, and specificity of 80%, whereas retinal specialists had similar sensitivity but only half of the specificity [29].

At the global eye health setting where screening expertise is limited (e.g., India and Thailand), DL tools were shown to be a great alternative, or even more superior, to human graders to screen for DR. In Thailand, Ruamviboonsuk et al. reported promising sensitivity (97%) and

Table 2 The summary of all the artificial intelligence systems with the respective training datasets and diagnostic performance for different retinal diseases using fundus photographs

AI systems	Year	Disease	Imaging modality	Race	Clinical validation	Independent testing datasets (retinal images)	AUC	Sensitivity	Specificity
Diabetic retinopathy									
Abramoff et al. [23]	2016	Referable DR (worse than mild DR)	Fundus photo	White	Messidor-2	874	0.980	96.8%	87.0%
Gulshan et al. [4]	2016	Referable DR	Fundus photo	White	EyePACS-1	9963	0.991	97.5%	93.4%
Gargya and Leng [6]	2017	Referable DR	Fundus photo	White	Messidor-2	1748	0.940	96.1%	93.9%
Ting et al. [5]	2017	Referable DR	Fundus photo	White	E-Ophtha	—	0.990	—	—
				Asians (Chinese, Malays, Indians, and others)	SIDRP 14-15	35,948	0.940	90.5%	91.6%
				Chinese	Guangdong	15,798	0.949	98.7%	81.6%
				Malay	SIMES	3052	0.889	97.1%	82.0%
				Indians	SINDI	4512	0.917	99.3%	73.3%
				Chinese	SCES	1936	0.919	100%	76.3%
				Chinese	BES	1052	0.929	94.4%	88.5%
				African	AFEDS	1968	0.980	98.8%	86.5%
				White	RVEEH	2302	0.983	98.9%	92.2%
				Hispanics	Mexican	1172	0.950	91.8%	84.8%
				Chinese	CUHK	1254	0.948	99.3%	83.1%
				Chinese	HKU	7706	0.964	100%	81.3%
Krause et al. [25]	2018	Referable DR	Fundus photo	White	EyePACS-2*	—	0.986	97.1%	92.3%*
Abramoff et al. [26]	2018	Referable DR (worse than mild DR)	Fundus photo	White	FDA Pivotal Trial	892	—	87.2%	90.7%

Li et al. [24]	2018	Referable DR	Fundus photo	Chinese	ZhongShan NIEHS	8000	0.989	97.0%	91.4%
				SIMES	15,679	0.962	93.9%	97.6%	98.5%
Ruanviboonsuk et al. [30]	2019	Referable DR	Fundus photo	Thai	Thailand Diabetes Registry	12,341	0.969	94.6%	99.2%
Gulshan et al. [31]	2019	Referable DR	Fundus photo	Indian	Sankara Aravind	25,326	0.987	96.8%	95.6%
Varadarajan et al. [29]	2020	DME	Fundus photo	Thai	Thailand dataset	3779	0.980	92.1%	95.2%
Bhuiyan et al. [28]	2020	Referable DR	Fundus photo	White	Kaggle dataset MESSIDOR dataset	1983	0.963	88.9%	92.2%
Bora et al. [33]	2021	Predict the progression to mild-or-worse DR	Fundus photo	White Thai	EyePACS Thai national diabetic patients registry	1033	0.890	85.0%	80.0%
Glaucoma suspect						974	—	82.6%	93.7%
Li et al. [7]	2018	CDR > 0.7 and glaucomatous changes	Fundus photo	Chinese	LabelMe	7976	0.79	0.77 (3-filed model) 0.70 (1-filed model)	
Ting et al. [5]	2017	CDR > 0.8 and glaucomatous changes	Fundus photo	Chinese, Malay, Indian and others	SIDRP 14-15	71,896	0.942	96.40%	87.20%
Shibata et al. [38]	2018	Glaucoma	Fundus photo	Japanese	Matsue Red Cross Hospital	110	0.965	NR	NR
Masumoto et al. [39]	2018	Glaucoma	Wide-field fundus photo	Japanese	Tsukazaki Hospital	282	0.872	81.30%	80.20%

(continued)

Table 2 (continued)

AI systems	Year	Disease	Imaging modality	Race	Clinical validation	Independent testing datasets (retinal images)		
						AUC	Sensitivity	Specificity
Phene et al. [37]	2019	Referable GON	Fundus photo	White, Indian	A: EyePACS B: Atlanta Veterans Affairs Eye Clinic Diabetic teleretinal screening program C: Dr.Shroff's Charity Eye Hospital, India	1205 9642 346	0.945 0.855 0.881	80.0% — —
Ko et al. [40]	2020	GON	Fundus photo	Chinese	Tapei Veterans General Hospital Drishti-GS	187 51	0.992 0.937	95.7% 73.6% (after fine-tuning)
Age-related macular degeneration (AMD)								
Burlina et al. [8]	2017	Referable AMD	Fundus photo	White	AREDS 1	26,764 images (AREDS 2)	0.94-0.96	71.00-88.40%
Burlina et al. [46]	2018	AMD Prediction	Fundus photo	White	AREDS 2	8088 images	1. Weighted K score (0.74-0.77); 2. overall mean estimation error for the 5-year risk	91.40-94.10%
Ting et al. [5]	2017	Referable AMD	Fundus photo	Chinese, Malay, Indian and others	SiDRP 14-15	71,896	0.931	93.20%
Grassmann et al. [9]	2018	Any AMD	Fundus photo	White	AREDS 1 Kora	33,886 5555	— —	100% (Late Stage AMD) 96.5% (Late Stage AMD)

Peng et al. [47]	2019	AREDS classifications	Fundus photo	White	AREDS	900	Accuracy: 67.1%
		Large Drusens					Accuracy: 74.2%
		Pigmentary abnormalities					Accuracy: 89.0%
		Late AMD					Accuracy: 96.7%
Keenan et al. [48]	2019	GA	Fundus photo	White	AREDS	59,812 with fivefold cross-validation	0.933–0.976 69.2% 0.939–0.976 76.3%
		Central GA					97.8% 97.1%
Liefers et al. [49]	2020	GA	Fundus photo	White	Rotterdam Study, Blue Mountains Eye Study	409 with fivefold cross-validation	Dice coefficient of 0.72 ± 0.26
Retinopathy of prematurity (ROP)							
Brown et al. [10]	2018	ROP plus disease	Retcam photo	White	i-ROP	100	—
		ROP pre-plus or worse					93.0% 100%
Tan et al. [57]	2019	ROP plus disease	Retcam photo	Australasian	Auckland Regional Telementicine ROP	1395	0.993 96.6%
					External validation	90	0.977 93.9%
Papilledema and optic disc abnormalities							
Ahn et al. [63]	2019	Pseudopapilledema	Fundus photo	Korean	Kim's Eye Hospital	1369	0.992 Accuracy 95.9%
Milea et al. [12]	2020	Papilledema	Fundus photo	Multiple ethnicities	19 clinical sites	1505	0.960 96.4%
Biousse et al. [62]	2020	Papilledema	Fundus photo	Multiple ethnicities	19 clinical sites	800	0.960 83.1% 94.3%

specificity (96%) of the DL system in a national screening program, with AUC of 0.99 [30]. In India, the DL system achieved a sensitivity and specificity of 89% and 92%, respectively (AUC of 0.96), in Aravind Eye Hospital, and 92% and 95%, respectively (AUC of 0.98), in Sankara Nethralaya Hospital [31]. In Africa, Bellemo et al. reported promising sensitivity (92%) and specificity (89%), with AUC of 0.97 for the diagnostic performance of a DL system in Zambia, a low middle-income country with the ratio of 3 ophthalmologists to 1000,000 population [32]. Although DL technology may have a great potential to help improve the access to DR screening for the under-resourced countries, the supporting infrastructure and availability of the tertiary care services are particularly crucial to support such technology.

For epidemiology, a DL system was also reported to accurately estimate the prevalence of DR and related systemic cardiovascular risk factors, as compared to the human graders, using a much shorter time frame (1 month vs 2 years) [14]. These findings suggest the potential role of DL systems to serve as the grading tool for many large-scale epidemiology or population-based studies. In addition, DL has been applied for the prediction of DR progression. Using color fundus photographs from 575,431 eyes with no DR, Bora et al. from Google Health created a DL system to predict the progression to mild or worse DR within 2 years, which achieved AUC of 0.79 and 0.78 on internal validation dataset for the three-field and one-field model, respectively. This DL system retained high prognostic value after adjusting for available risk factors, such as the HbA1c level, years with diabetes, and insulin use [33].

Of all the AI algorithms in ophthalmology or medical fields, the DL systems for DR is probably at the closest stage for clinical translation. To achieve such success, it is critical to build a surrounding ecosystem to support the deployment of DR systems in the real-world practice, including IT systems and integration with the electronic health records, medical ethics, safety regulations, financial and reimbursement systems, tailoring to the respective healthcare system regulations, medical expertise, and patients' demographics.

Glaucoma

Glaucoma is a chronic progressive optic neuropathy with characteristic visual field (VF) loss associated with intraocular pressure elevation, leading to irreversible vision loss without early diagnosis and prompt treatment [34, 35]. The global prevalence of glaucoma is 3.4% for people aged 40–80 and it is projected to be approximately 112 million by 2040 [36]. Therefore, screening and monitoring for glaucoma are of paramount importance and AI applications might play a pivotal role in diagnosis and surveillance of this condition. The DL systems are broadly divided into the fundus photos, OCTs, and Humphrey VF-based algorithms in detecting glaucoma suspect, glaucoma, or glaucoma progression. This section will focus on the fundus-based glaucoma AI algorithms.

For glaucoma suspect detection, two DL systems, tested on >10,000 retinal images, reported to have AUC of >0.90, sensitivity >90%, and specificity >80%, [5, 7] although the definition of cup-to-disc ratio was slightly different between the two studies (Li et al. -0.7 or worse; Ting et al. 0.8 or worse). Tested with three independent datasets, the DL algorithm developed by Phene et al. using fundus images alone achieved AUC for referable glaucoma of 0.945, 0.855, and 0.881 in dataset A, B, and C, respectively [37]. The decrease in performance with datasets B and C was attributed to the differences in reference standard, in which the graders had access to additional clinical data, such as a full glaucoma workup. Using smaller testing datasets (<500 fundus photos), Shibata et al., [38] Masumoto et al., [39] and Ko et al. [40] reported AUC >0.85 in confirming the diagnosis of glaucoma, using the diagnosis made on clinical examination and VF.

The development of glaucoma AI algorithms is often challenged by the inconsistent ground truth, and lack of consensus on the definition of glaucoma suspect or definite glaucoma. In addition, the diagnosis of glaucoma requires not only multimodal imaging, but also time sequence to confirm the glaucoma progression. Thus, it is hard to conduct a head-to-head comparison between different glaucoma AI algorithms, as each of them may be built to answer different research questions. While the optic disc fundus imaging of the retina is the least expensive imaging modality to

conduct structural assessment of the optic nerve, glaucoma detection consists of combined structural and functional assessments, using more sophisticated imaging devices, such as OCT [41–43]. Further details will be discussed in the latter section.

Age-Related Macular Degeneration

By 2040, 288 million patients may have some forms of AMD and approximately 10% may be affected by intermediate AMD or worse [44]. With aging population, DL algorithms could be deployed as alternative tools to aid diagnosis and disease surveillance, tackling the urgent clinical need to screen these patients for further evaluation in healthcare centers. Similar to DR and glaucoma, most DL algorithms have reported robust diagnostic performance in detecting AMD.

The majority of the AMD algorithms reported in the literature were developed using the Age-Related Eye Disease Study (AREDS) with different clinical questions, given that it is available to public request with the AREDS committee approval [45]. Using the best training model, Burlina et al. reported excellent diagnostic accuracy and AUC (both >90%) in detection of referable AMD (defined as intermediate AMD or worse based on AREDS classification), [8] and accuracy of >80% in estimating 5-year risk of developing advanced AMD [46]. Using different technical methods, Grassmann et al. [9] reported 100% sensitivity and 96.5% specificity in detecting late AMD, although the performance in detection of early AMD is suboptimal. Peng et al. [47] reported that the DL system could perform better than the retinal specialists to determine the AREDS grading (accuracy 67% vs 60%), and specific AMD lesions (e.g., large drusen and pigmentary abnormalities) (AUCs of 0.94 vs 0.93). Groups studying the automated detection of geographic atrophy (GA) also demonstrated high accuracy and noninferiority to human graders [48, 49]. Using multicenter datasets, Liefers et al. [49] developed a DL model for automatic segmentations of GA, which identified nine automatically calculated structural biomarkers being significantly correlated with GA growth rate.

Besides detection of AMD from fundus images, AI has been applied for predicting risk

of progression to late AMD. Using fundus photos and demographic data from over 1800 AREDS subjects, Bhuiyan et al. developed a logistic model tree machine learning (ML) algorithm to predict the risk of an individual with early or intermediate AMD progressing to late AMD within 1 or 2 years. By combining the AMD 12 level severity score and sociodemographic clinical data, the ML model could predict the risk of AMD progression with accuracy of 86.36%, sensitivity of 92.42%, specificity of 84.39% for 2-year incident late AMD, and similar performance for predicting the incidence of late AMD in 1 year [50].

At present, the application of fundus-based AMD algorithms is yet to penetrate the clinical space due to the following reasons. First, although the American Academy of Ophthalmology (AAO) guidelines recommend regular follow-up once every 2 years for patients with intermediate AMD or worse, not many countries worldwide are currently practicing this exercise, given the resultant of unnecessary referrals to the tertiary eye care settings. Second, there is no effective treatment to delay or prevent progression to advanced AMD, although various interventions were attempted (e.g., AREDS2 formulation and smoking cessation). Third, to accurately diagnose active choroidal neovascularization, it is more accurate to perform OCT scans for diagnosis and disease monitoring during the treatment course [15, 51, 52]. More details are discussed in the section under OCT AI algorithms.

Retinopathy of Prematurity

ROP is characterized by the growth of abnormal fibrovascular structures at the junction between vascularized and avascular peripheral retina. Globally, approximately 19 million children are estimated to suffer from visual impairment, [53] with ROP accounting for up to 18% of childhood blindness [54]. Given that ROP clinical examination can be technically challenging, many countries, unfortunately, do not have such expertise to perform ROP screening. In the USA, a DL system for ROP was developed using approximately 5500 Retcam retinal photographs collected over a 5-year period from eight US academic institutions for training and testing [10]. Compared to

the ROP clinical experts, the DL system demonstrated robust sensitivity (>90%) and specificity (>90%) in detection of plus and pre-plus diseases, [10] and excellent AUCs (>0.90) in detecting type 1 ROP and “clinically significant ROP” that warrant an urgent referral to a specialty center [55]. This DL system was further evaluated retrospectively using an operational ROP telemedicine program. With limited sample size (81 infants, 613 eye encounter images), Greenwald et al. reported that the AUC of the vascular severity score for the detection of referral-requiring ROP was 0.99, with 100% sensitivity and 90% specificity [56]. In New Zealand, using 6974 Retcam retinal images from infants, a DL algorithm was developed to detect plus disease with AUC of 0.993, sensitivity of 96.6%, and specificity of 98.0% [57].

Future research is of great value to test these ROP systems in different population, in order to increase the generalizability of such system. In the developed countries where there are established ROP services, the bar of having AI algorithms replacing clinicians in ROP screening may be reasonably high, provided that there are always potential occasions where the DL systems could misdiagnose or underdiagnose ROP cases that may carry clinical, medico-legal, and psychosocial implications. On the other hand, although the ROP DL systems are extremely useful in countries without existing ROP screening services, it may still be challenging to deploy such technically demanding system as the Retcam imaging. This, again, highlights the need of having a robust healthcare ecosystem to support the deployment of any AI algorithm in clinical care.

Papilledema and Optic Disc Abnormalities

Examination of the optic nerve head is an important component of the clinical examination when indicated, but general physicians and nonophthalmic specialists may lack confidence in using direct ophthalmoscopy, especially without pharmacologic pupillary dilation [58, 59]. Alternative use of nonmydriatic ocular fundus photography has been adapted in the emergency departments and outpatient clinics [58, 60,

61]. However, these fundus images need to be interpreted on-site by physicians or remotely by ophthalmologists via telemedicine, which may not be always feasible. Several groups have evaluated the performance of AI systems in detecting papilledema and optic disc abnormalities from fundus photographs. Using 15,846 fundus images from multiple ethnic groups in 11 countries, Milea et al. developed a DL classifier to discriminate papilledema from normal discs and other disc abnormalities (e.g., anterior ischemic and inflammatory optic neuropathies, optic atrophy), achieving AUC of 0.990, sensitivity of 93.5%, and specificity of 96.2% [12]. Similarly, the Brain and Optic nerve Study with Artificial Intelligence (BONSAI) DL system, which was developed with multiethnic and multicenter datasets, demonstrated similar or even superior performance in classifying optic disc appearance when comparing to neuro-ophthalmologists. For classification of papilledema, the BONSAI system achieved AUC of 0.96, sensitivity of 91.5%, and specificity of 94.3% [62]. Using a smaller dataset, Ahn et al. reported that the ML classifier could differentiate pseudopapilledema from normal discs and discs with other abnormalities with an accuracy of 95.89% and AUC of 0.992 [63]. Introducing these DL models to the routines in emergency departments and outpatient clinics does not replace human experts, instead it may facilitate effective triage and management of patients, in particular when experts are not available.

Systemic Diseases

Besides retinal diseases, a number of DL algorithms have been developed to detect systemic diseases from fundus photographs, with cardiovascular disease (CVD) being the mostly studied. CVD is a major cause of morbidity and mortality globally [64]. It is now recognized that microvascular pathology plays a key role in processes leading to the development of subclinical vascular disease and clinical CVD events [65–67]. In the past, physicians perform a fundus examination in patients with hypertension and follow classification system to determine the presence and severity of retinal vascular damage, such as retinal arteriolar narrowing, arteriovenous nicking, retinal

hemorrhages, and cotton-wool spots, as a mean to estimate CVD risk [68–73]. Together with left ventricular hypertrophy and renal impairment, most international hypertension management guidelines include the retinal vascular changes (or hypertensive retinopathy) as an indicator of target organ damage, and that its presence should be an indication for a more aggressive approach in managing these hypertensive patients [70–73].

Broadly speaking, there are three basic approaches to predicting systemic diseases from a retinal image (Table 3). First, DL algorithms could be applied for estimating systemic risk factors based on fundus photographs. Using two large databases (48,101 patients from the UK Biobank and 236,234 patients from EyePACS), Poplin et al. have developed DL models using retinal fundus images to predict multiple CVD risk factors simultaneously, including age, gender, smoking status, blood pressure, body mass index, HbA_{1c}, and major cardiac events [13]. The reported AUC of the DL model (AUC = 0.70) for predicting CVD events was comparable to that of the composite European SCORE risk calculator (AUC = 0.72). In a further validation using an Asian dataset, the performance of the DL models was similar [74]. A surprising finding from this study was the ability of the DL algorithm to accurately predict the gender of a patient from fundus image, with AUC of 0.97, and this has certainly generated widespread discussion among the ophthalmology community. In addition, Rim et al. evaluated 47 DL algorithms in predicting 47 systemic biomarkers based on fundus photographs, with use of 236,257 images from diverse population datasets. The DL systems were shown to quantify body composition indices (muscle mass, height, and bodyweight) and creatinine from fundus photos alone, while 37 biomarkers could not be predicted well from retinal photographs ($R^2 \leq 0.14$ across all external test sets) [75].

Second, DL models have been developed for the purpose of replacing another biomarker with findings from retinal images. Cheung et al. evaluated the correlation between retinal-vessel caliber and CVD risk factors, using multiethnic multi-country datasets of over 70,000 images to develop a DL system, the Singapore I vessel assessment-

DL system (SIVA-DLS). [76] High agreement was achieved by the SIVA-DLS and human graders in measuring central retinal artery equivalent (CRAE) and central retinal vein equivalent (CRVE), with overall interclass correlation coefficients of 0.82–0.95. Figure 3 demonstrates the prediction of retinal-vessel caliber by the SIVA-DLS. The model showed superior or comparable performance than expert graders in identifying the relationships between measurements of retinal-vessel caliber and CVD risk factors. Furthermore, the SIVA-DLS measured retinal-vessel caliber was associated with incident CVD events. Using 15,408 fundus images, Chang et al. built a DL model to predict carotid artery atherosclerosis, named deep-learning funduscopic atherosclerosis score (DL-FAS), with an AUC, sensitivity, and specificity of 0.713, 58.3%, and 89.1%, respectively. The DL-FAS was used as the main exposure to follow patients for the primary outcome of death. They found that participants with DL-FAS greater than 0.66 had higher risk of CVD deaths compared to those with DL-FAS less than 0.33 (hazard ratio: 8.33) [77]. To estimate coronary artery calcium score (CACS) from fundus photographs, Son et al. built a DL algorithm using 44,184 images, which could recognize subjects with CACS over 100 from those of no CAC, with an AUC of 0.823 for unilateral fundus images and 0.832 for bilateral ones. However, whether the CACS could predict clinical CVD events remains to be explored [78].

Finally, DL techniques could be used for detecting specific diseases based on fundus photographs, such as hypertension, hyperglycemia, dyslipidemia, anemia, and chronic kidney disease. With use of over 2000 vessel-segmented retinal photographs, Dai et al. developed a DL model to generate heat maps for predicting hypertension, with an AUC of 0.651, an accuracy of 60.94%, and a specificity of 51.54% [79]. Another DL framework built by Zhang et al. using 1222 fundus images was shown to predict hypertension, hyperglycemia, and dyslipidemia with an AUC of 0.766, 0.880, and 0.703, respectively [80]. Mitani et al. developed ML algorithms that could detect anemia from retinal fundus photographs alone with an AUC

Table 3 Three basic approaches to predicting systemic diseases from a retinal image and the relevant studies

1. AI-DL algorithms to estimate systemic risk factors from retinal fundus images							
AI systems	Year	Disease	Imaging modality	Race	Clinical validation	Independent testing datasets (retinal images)	Results
Poplin et al. [13]	2018	CVD	Fundus photo	White, Black, Asian	UK Biobank EyePACS-2 K	12,026 999	Age: mean absolute error within 3.26 years Gender: AUC = 0.970 Smoking status: AUC = 0.710 Systolic blood pressure: mean absolute error within 11.23 mmHg Major adverse cardiac events: AUC = 0.70
Rim et al. [75]	2020	47 Systemic biomarkers	Fundus photo	Asian, European	Severance Gangnam Hospital Beijing Eye Study SEED study UK Biobank	4662 4234 63,275 50,732	Quantification of muscle mass ($R^2 = 0.52$), height ($R^2 = 0.42$), and bodyweight ($R^2 = 0.36$), creatinine ($R^2 = 0.38$) 37 could not be predicted well from retinal photographs ($R^2 \leq 0.14$ across all external test sets)
2. AI-DL algorithms to replace another biomarker with retinal image for CVD prediction							
AI systems	Year	Purpose	Imaging modality	Race	Clinical validation	Independent testing datasets (retinal images)	Results
Chang et al. [77]	2020	DL-fundoscopic atherosclerosis score to predict carotid artery atherosclerosis	Fundus photo	Korean	Health Promotion Center of Seoul National University Hospital	634 patients	AUC 0.713 Sensitivity 58.3% Specificity 89.1%
Cheung et al. [76]	2020	Measure retinal-vessel diameter to assess CVD risk	Fundus photo	Multi-ethnic	21 multi-ethnic, multi-country datasets	64,827	Retinal-vessel calibre measured by SIVA-DLS model and human has interclass correlation of coefficients of 0.82–0.95 Retinal-vessel calibre measured by SIVA-DLS is associated with CVD risk factors and incident CVD

3. AI-DL algorithms to detect specific diseases from retinal fundus images						
AI systems	Year	Disease	Imaging modality	Race	Clinical validation	Independent testing datasets (retinal images)
Dai et al. [79]	2020	Hypertension	Fundus photo	Chinese	Shenyang He Eye Hospital	2012
Zhang et al. [80]	2020	Hypertension Hyperglycemia Dyslipidemia	Fundus photo	Chinese	Xinxiang county, Henan	1222
Mitani et al. [81]	2019	Anemia	Fundus photo	White	UK Biobank	11,388
Sabanayagam et al. [82]	2020	Chronic kidney disease	Fundus photo	Chinese, Indian, Malay	SEED BES SP2	1297 1538 3735
Son et al. [78]	2020	Predict coronary artery calcium score	Fundus photo	Korean	Seoul National University Bundang Hospital	44,184 0.823
						AUC — — —

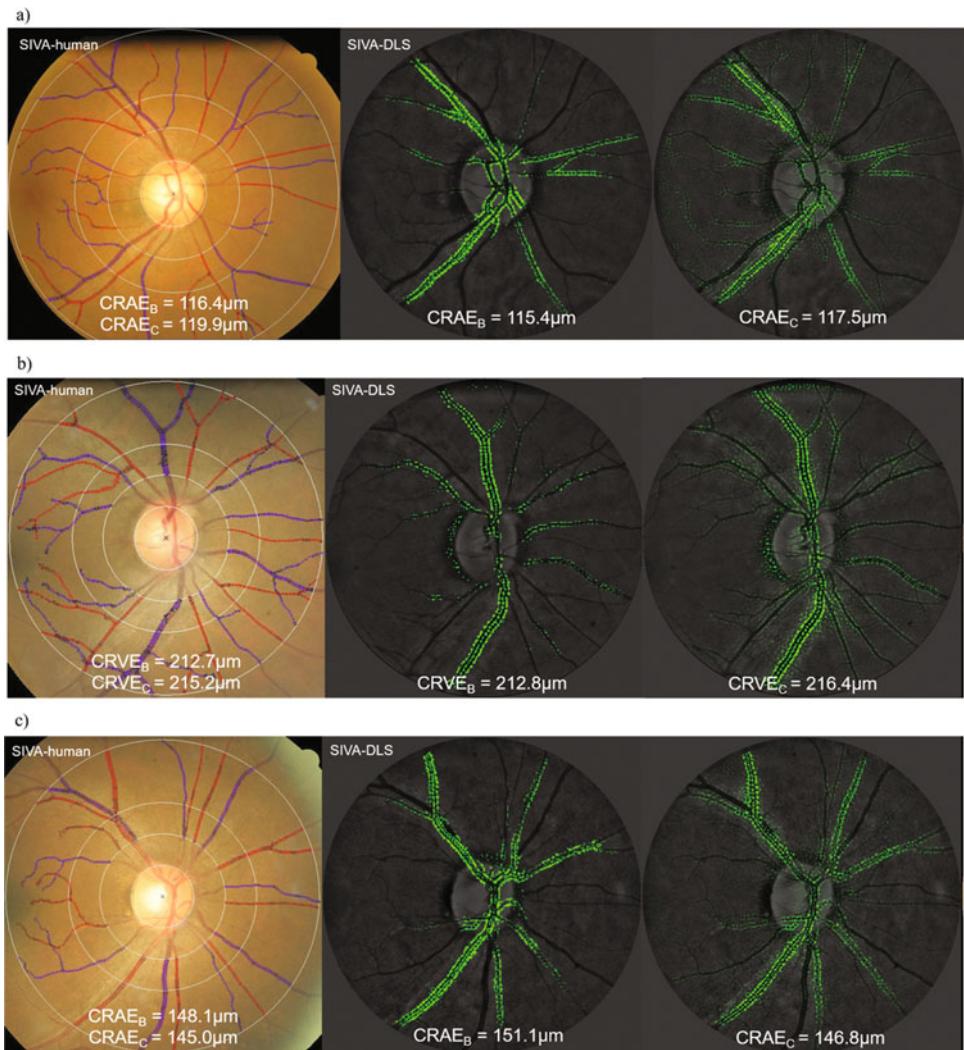


Fig. 3 An example of CRAE and CRVE prediction using SIVA-human and SIVA-DLS. The retinal-vessel calibers were calculated at two regions: one from 0.5 to 1.0 disc

diameter away from disc margin (zone B: CRAE_B and CRVE_B) and one from 0.5 to 2.0 disc diameters away from the disc margin (zone C: CRAE_C and CRVE_C) [76]

of 0.74. The ML model achieved higher performance in study participants with self-reported diabetes, as evidenced by an AUC of 0.89 in predicting anemia [81]. Sabanayagam et al. built a DL framework using fundus images of over 5000 patients from multiethnic, population-based datasets, which demonstrated good performance in predicting chronic kidney disease from fundus photographs, with an AUC of 0.911 in the validation dataset, and 0.733 and 0.835 in two external testing datasets [82].

Although it is interesting to predict systemic diseases with DL using retinal imaging, the clinical relevance and impact is yet to be defined. For example, to impact patients' management outcome, future research is of great value to explore using the retinal imaging CVD scores to stratify patients with CVD into high, moderate, or low risks, in addition to all the existing CVD risk calculator parameters, and this will require close collaboration between ophthalmologists and cardiologists in development of such tools.

Optical Coherence Tomography

OCT, particularly high-resolution spectral domain OCT (SD-OCT), is now the dominant imaging modality in ophthalmology, with more than 30 million OCT scans being acquired per year worldwide – more than all other modalities combined [83]. OCT is of particular value in the diagnosis and management of retinal diseases such as CNV in neovascular AMD and DME, the commonest causes of irreversible vision loss in many countries [84, 85]. In fact, the use of OCT to allow personalized treatment of retinal diseases has allowed cost savings estimated at \$9 billion, a 21-fold return on initial US government investment in developing this technology [86].

To date, the OCT-based DL systems focus on three major areas, namely the segmentation, classification, and predictive tasks. In segmentation tasks, DL has been adopted to segment “fluid” within the retina, both intraretinal fluid (IRF) and subretinal fluid (SRF). The automated quantification of these fluids could be used to guide anti-vascular endothelial growth factor (VEGF) treatments in retinal diseases [87, 88]. Using the OCT data of 1127 eyes from the AREDS2 10-year follow-on study, Keenan et al. reported that the Notal OCT Analyzer (NOA), a ML classifier, could detect retinal fluid with higher accuracy and in particular higher sensitivity than expert graders (accuracy of 0.851 vs 0.805, sensitivity of 0.822 vs 0.468, and specificity of 0.865 vs 0.970) [89]. A novel, alternative approach has recently been suggested for detecting early AMD biomarkers, via annotation of OCT scans for biomarkers at the B-scan rather than pixel level, with subsequent training of a DL system to classify the scans. Class activation maps can then be used on each scan to provide approximate delineation of the biomarkers in a quick and entirely automated manner [90]. The classification and predictive tasks will be discussed in the following section.

DL Algorithms for Retinal Diseases Using Macula-Centered OCT

Using close to 100,000 OCT scans, Lee et al. reported excellent AUC and diagnostic accuracy

(>90%) in discriminating “normal” versus neovascular AMD patients (Table 4) [51]. Kermany et al. subsequently reported a robust diagnostic performance in discriminating DME, drusen, neovascular AMD, normal, with an accuracy of 96.6%, a sensitivity of 97.8%, and a specificity of 97.4%. In addition, this study also showed that transfer learning could potentially reduce the need of having large training datasets to train a robust AI algorithm [15]. Developed with OCT scans from a Korean dataset, Rim et al. demonstrated that the DL model could be generalized onto an external validation dataset of 91,509 OCT scans from American population, with AUC of 0.952 at individual level [91]. De Fauw et al. developed a two-stage DL framework by decoupling the segmentation and classification network, using 14,884 OCT scans. The DL system was able to detect those who require urgent referrals with excellent performance (AUC of >0.90), using two different OCT systems (Topcon and Spectralis) [16]. This DL algorithm utilized nine contiguous OCT scans, a three-dimensional U-net architecture, and intermediate tissue representation to output-automated segmentations across 15 different label classes. These labels encompass a range of novel OCT biomarkers, including three forms of pigment epithelial detachment (PED: fibrovascular, serous, and drusenoid) and sub-retinal hyperreflective material. The authors also highlighted the need to perform domain adaptation to fine-tune the DL algorithm for a new device, which brought down the total error rate for referral suggestions from 46.6% to 3.4%.

DL has also been shown to be useful in assisting the management of neovascular AMD with OCT measurements. Schmidt-Erfurth et al. applied a DL framework for automated detection and quantification of IRF, SRF, and PED before and after anti-VEGF treatment, which revealed that higher dose and more frequent dosing correlated with least residual fluid volumes [92]. DL model may become a more reliable and reproducible tool to quantify the response to anti-VEGF treatment, comparing to the qualitative assessment currently used in clinical practice. For the predictive tasks, Schmidt-Erfurth et al. used the HARBOR data to

Table 4 The summary of all the artificial intelligence systems with the respective training datasets and diagnostic performance for different retinal diseases using optical coherence tomographs

AI systems	Year	Disease	Imaging modality	Race	Clinical validation	Independent testing datasets (retinal images)		AUC	Sensitivity	Specificity
Macula OCT										
Lee et al. [51]	2017	Exudative AMD	Spectralis OCT	White	Clinic-based	20,613	0.928	84.60%	91.50%	
Treder et al. [52]	2018	Exudative AMD	Spectralis OCT	White	Clinic-based	100	NR	92%	96%	
Kermer et al. [15]	2018	Multi-class comparison (CNV, DME, Drusen and normal)	Spectralis OCT	White	Clinic-based	1000	0.99	97.80%	97.40%	
De Fauw et al. [16]	2018	Urgent, semi-urgent, routine, and observation only	Topcon OCT	White	Clinic-based	997	0.992	Accuracy: 94.5% (Urgent referral)		
Schmidt-Erfurth et al. [93]	2018	AMD (Prediction of visual acuity)	Spectralis OCT	White	Harbor Clinical trial	116	0.999	Accuracy: 96.6% (Urgent referral)		
Asoaka et al. [101]	2019	Glaucoma	RS3000, RS (SD-OCT)	Japanese	Tokyo University Hospital, Tajimi Iwase Eye Clinic	196	0.937	82.5%	93.90%	
Rim et al. [91]	2020	Neovascular AMD	Spectralis OCT	Korean American	Severance Hospital	493 participants	0.999	—	—	
Optic nerve OCT										
Ran et al. [103]	2019	GON	Cirrus OCT	Chinese	Hong Kong Eye Hospital	976 (3-D) 976 (2-D)	0.969 0.921	89% 85%	96% 85%	

		Chinese	Prince of Wales Hospital	546	0.893 (3-D)	79%	84%
		Chinese	Tuen Mun Eye Center	267	0.770 (2-D)	72%	75%
		American	Byers Eye Institute	1231	0.897 (3-D)	90%	79%
					0.752 (2-D)	78%	64%
Medeiros et al. [98]	2019	GON	Optic disc photographs and SD-OCT	White	Duke Glaucoma Repository	0.917 (3-D)	78%
Kim et al. [102]	2020	Glaucoma	Cirrus SD-OCT	Korean	Samsung Medical Center, Kangbuk Samsung Hospital	0.981 1700 (internal validation) 1420 (external validation)	86% 97.8% 98.4%

develop ML models to predict visual acuity (VA) in patients receiving ranibizumab for neovascular AMD [93]. They applied automated segmentation algorithms (using both graph-based and DL approaches) to the OCT scans, allowing segmentation of total retinal thickness, IRF, SRF, and PED, using random forest regression to predict VA at baseline and at 12 months. In the study, they found that the patients with good VA at baseline, and then at each follow-up for 3 months, were likely to have good VA at 12 months. In addition to the prediction of treatment results, Yim et al. built a DL system to predict the progression to exudative AMD in fellow eyes of patients with exudative AMD in the other eye, by combining 3D OCT images and corresponding automatic tissue maps, which outperformed five out of six experts. This DL system achieved sensitivity of 80% at specificity of 55% and sensitivity of 34% at specificity of 90% in predicting the conversion to exudative AMD within 6 months when compared with the ground truth of “conversion scan” rather than injection, corresponding to false positives in 9.6% of scans at the 90% specificity point [94]. These risk stratification and prediction models may have significant potential in the management of AMD by critically identifying the point of conversion from early or intermediate to exudative AMD, allowing timely follow-up and treatment.

A recent study adapted multimodal retinal image analysis consisting of fundus photographs, OCT, and OCT angiography scans, for the detection of intermediate AMD. Despite a relatively small training dataset of 75 participants, Vaghefi et al. demonstrated that the DL system accuracy increased from 91% to 96% in detecting intermediate AMD by combining multiple modalities, comparing to using OCT alone [95]. It is therefore crucial for the future research to evaluate the generalizability of these DL systems in a larger international multiethnic cohort, incorporating multimodal approach with clinical data, fundus photographs, and OCT scans. Apart from screening purposes, it will be of great value to generate new algorithms to predict and prognosticate the functional, structural, and treatment outcome for AMD patients, with appropriate stratification of the risk profiles.

DL Algorithms for Glaucoma Using Optic Disc-Centered OCT

Glaucomatous optic neuropathy (GON) is characterized by thinning of the retinal nerve fiber layer (RNFL) and optic disc cupping, in connection with loss of retinal ganglion cells (RGCs). Assessment of peripapillary RNFL thickness by OCT is now commonly used to detect GON for diagnosis and management of glaucoma, especially mild to moderate glaucoma. OCT has also been proposed to be used for screening GON in high-risk communities (e.g., elderly people, people with high myopia, people with a family history of glaucoma, or people with high intraocular pressure) [96, 97]. However, experienced glaucoma specialist or highly trained assessors are required to interpret the OCT results. Thus, AI algorithms may play an important part in detecting and managing GON.

Besides fundus-based glaucoma AI algorithms as discussed earlier, several groups have explored the use of OCT measurements for the training of a DL system to quantify GON on 2D fundus images [98, 99]. Using a total of 32,820 pairs of optic disc photographs and OCT RNFL scans, Medeiros et al. showed that the prediction of RNFL thickness from optic disc photos are similar to the OCT findings (83.3 μm vs 82.5 μm , $p = 0.164$). The AUC in differentiating GON from health eyes were 0.944 [98]. In addition, a DL model recently reported by Medeiros et al. demonstrated the ability to discriminate glaucoma progressors from nonprogressors, based on quantitative estimates of RNFL changes from fundus photographs, with an AUC of 0.86 [100].

Several DL frameworks have been developed to detect glaucoma using OCT measurements such as the RNFL thickness and ganglion cell complex (GCC) images [101, 102]. Asaoka et al. built a classifier by first pretraining with 4316 OCT images and then further training with 178 images obtained with a different OCT instrument. This transform model achieved AUC of 0.937, sensitivity of 82.5%, and specificity of 93.9% in differentiating early glaucoma eyes from normal eyes [101].

Subsequently, Ran et al. developed a three-dimensional (3D) DL system from 4,877 volumetric OCT scans for automated detection of GON, aiming to utilize the GON features of 3D

volumetric OCT cube that are not shown in 2D fundus photographs [103]. In the primary validation, they showed that the 3D DL algorithm outperformed the 2D DL algorithm trained on fundus images (AUCs 0.969 vs 0.921, $p < 0.001$). In the external validation, the 3D DL algorithm was tested in three other independent datasets and showed good performance, with AUCs of 0.893–0.897, sensitivities of 78–90%, and specificities of 79–86%. This study suggests that screening with the 3D DL system is much faster than conventional glaucoma screening methods (i.e., by experienced specialists), could be done automatically, and does not require a large number of trained personnel on site. Future work such as comparison of discriminative performance between RNFL map and the 3D DL system, and testing the ability of the 3D DL system to detect GON damage in people with suspected glaucoma are warranted [104].

DL Algorithms for Glaucoma Using Anterior-Segment OCT

Anterior-segment OCT (AS-OCT) has been used increasingly to assess the anterior chamber angle for differentiating open-angle glaucoma, closed-angle glaucoma, and angle closure diseases. Imaging with AS-OCT has several advantages over conventional gonioscopic examination, such as noncontact approach and better reproducibility [105, 106]. Xu et al. developed a DL model with multiclass CNNs to classify the anterior chamber angle into open- or closed-angle using 4036 AS-OCT images collected from the Chinese-American Eye Study. They found that using the ResNet-18 architecture, the DL algorithm can achieve an AUC of 0.933 on the cross-validation dataset and 0.928 on the testing dataset for detecting gonioscopic angle closure. Although inclusion of more anterior-segment OCT images for the classification and external validation are needed, the results of this study indicate that the detection of eyes with gonioscopic angle closure may be automated using DL framework with favorable performance compared to the manual and semi-automated methods [107]. In addition, the authors also suggested that DL algorithm may be further applied to identify patients with early angle closure diseases who are at high risk of

GON and may require laser peripheral iridotomy [107, 108].

Visual Fields

VF is a pivotal test in the diagnosis and monitoring of glaucoma patients. Compared to optic disc photographs or OCT images, VF data points usually have low dimensionality and high noise. Due to the lack of consensus in the definition of glaucoma, DL systems may help play a role in defining the minimal thresholds to diagnose glaucoma. Using the pattern deviation (PD) probability plots, Li et al. developed a CNN to discriminate glaucomatous from normal eyes, with 93.2% sensitivity and 82.6% specificity [109]. For identification of glaucoma progression, various ML and DL techniques have been attempted. First, Yousefi et al. used an alternative Gaussian mixture and expectation maximization method to decompose VFs along different axes to detect VF progression [110]. This approach was as good or superior to current algorithms, including glaucoma progression analysis, visual field index, and mean deviation slope, in detecting VF progression. Second, Garcia et al. reported that Kalman filtering (KF), a ML technique, is effective in forecasting disease trajectory in a total of 263 normal-tension glaucoma patients in prediction of 2-year mean deviation (MD) forecast than linear regression of MD [111]. Third, using approximately 32,000 Humphrey VFs with more than 1.7 million perimetry points, Wen et al. trained a CNN that is able to use a single Humphrey VF to generate predictions of future Humphrey VFs for up to 5.5 years, [112] with an excellent correlation of the MD between the predicted and actual Humphrey VF (average difference of 0.441 dB).

Several AI algorithms have been developed for relating structural changes of glaucoma on OCT to functional loss on VF [113–115]. Using over 25,000 pairs of standard automated perimetry (SAP) and SD-OCT measurements from patients with glaucoma or glaucoma suspect, Mariottini et al. built a CNN that is capable of predicting SAP sensitivity thresholds based on RNFL defects seen in SD-OCT, evidenced by a mean absolute error of 4.25 dB and an average correlation

coefficient of 0.60 ($P < 0.001$) with the measured values from SAP [113]. With a smaller training dataset, the DL model developed by Park et al. could predict the VF results from OCT images with a root mean square error of 4.79 ± 2.56 dB [115]. Application of these AI-based structure-function maps may assist the interpretation of OCT and VF results in clinical practice, especially for patients who may not be able to conduct VF assessment. Nevertheless, glaucoma is a disease affected by many clinical factors; future research is needed to combine the patients' demographic profile, intraocular pressure, and structural and functional tests in a multimodal approach to develop more accurate algorithms.

Infantile Facial Video Recording

In China, Long et al. reported a DL system that could discriminate the infantile behavioral dynamics of visual impairment (Fig. 4) [116]. In this study, a 5-minute video was recorded for 4196 infants, and analyzed for four main categories (13 subtypes) of behaviors labeled using a consensus definition: 1) eyeball movement (strabismus, nystagmus, and incongruous binocular movement); 2) hand-related behaviors (eye rubbing, pressing, and poking); 3) fixation-related behaviors (compulsive light gazing, compensatory head position, motionless fixation, and poor fixation); and 4) eyelid reaction (frequent blinking, squinting, and frowning), diagnosed independently by five experienced researchers and ophthalmologists, with the disagreement cases arbitrated by three senior professors. Based on the facial behaviors, the DL system was able to discriminate mild visual impairment vs healthy, severe vs mild visual impairment, and various ophthalmic diseases from healthy vision, with AUC of 0.852, 0.819, and 0.816–0.930, respectively. Given the convenience of a 5-minute video recording, this approach may be a useful alternative to identify and monitor infant with visual problems. Nevertheless, the generalizability of the algorithm may need to be further evaluated in the non-Chinese settings considering the differences in facial features. How to integrate the video recording in the busy clinical workflow is another

issue to be dealt with for the clinical application of the video-based AI algorithm.

Electronic Health Records

Electronic health records (EHR) have been adopted by the healthcare users since over a decade ago [117]. In 2018, DL using EHR has been shown to be effective in predicting the in-hospital mortality, 30-day unplanned readmission, and prolonged length of stay in two US academic centers [118]. In ophthalmology, Baxter et al. described using the structured EHR data and ML to predict the need for surgical intervention in 385 patients with primary open angle glaucoma [119]. Using the various statistical and ML models (e.g., multi-variable logistic regression, random forests, and artificial neural networks), the authors reported an AUC of 0.67 in discriminating patients with progressive disease requiring glaucoma surgeries. In this study, they also found that the use of certain ophthalmic and systemic medications (e.g., lipid lowering agents, non-opioid analgesics, macrolide antibiotics, and calcium blockers) were associated with decreased needs for glaucoma surgeries, suggesting the usefulness of some systemic data, in addition to eye-specific data, in the predictive modeling.

Although big data is always considered as a great resource to build data-based algorithms, the real-world data, unfortunately, are always sparse, lack organization and reference standards. In addition, the algorithm developed using a specific population data may not be generalizable to others, given the differences in patients' demographics, as well as epidemiological and geographical risk factors. Thus, it is important to ensure robust external independent testing, or the need for the specific algorithm to be fine-tuned further based on the specific population needs.

Image Quality Assessment

Image quality is of paramount importance for the training and performance of a DL algorithm. In the laboratory setting, preprocessing of images by

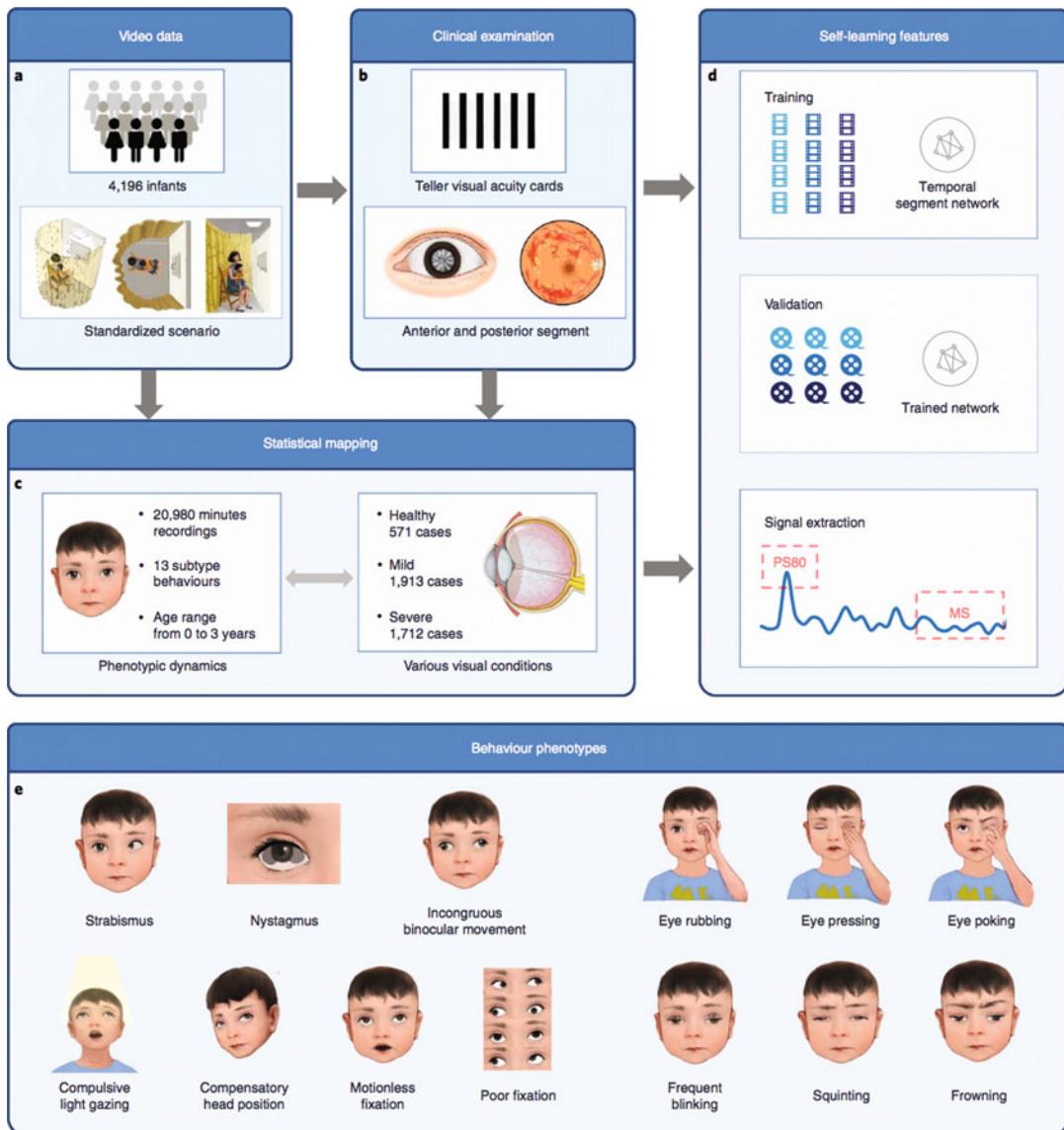


Fig. 4 Overall study pipeline. A. A total of 4196 infants' panoramic dynamics of behavioral phenotypes with various visual conditions were collected under a designed standardized scenario. B. Each infant underwent Teller visual acuity cards and structural examinations using a slit lamp and fundoscope. C. On the basis of the records, the statistical associations between behaviors and severities of visual impairment were calculated. D. A temporal segment

network was trained to self-learn features from the behavioral phenotypes in the training set (video symbols in the upper box). After testing the deep learning system performance using an independent validation dataset (film-reel symbols in the middle box), the trained network was used to extract two indices (MS and PS 80) to discriminate behaviors. E, Static representations of 13 behaviors [116]

filtering out those with poor quality is usually conducted during the development of DL algorithm, given that a suboptimal image may have impact on the diagnostic performance. In the real-world clinical setting, however, image quality could be compromised by several factors, such

as small pupil size, cataract, and patient's cooperation. It is therefore essential for a DL model to automatically assess the image quality with the gradability algorithm. Using 6139 retinal fundus images from preterm infants during routine ROP screenings, Coyner et al. developed a DL

framework for automated assessment of fundus images quality in ROP, where quality was defined as the ability to confidently assess an image for the presence of ROP. This DL model achieved comparable performance to that of human experts, with AUC of 0.965 for differentiating acceptable quality and not acceptable quality [120]. With use of over 40,000 ultra-widefield fundus images, the DL model built by Li et al. could filter out sub-optimal images with high sensitivity and specificity (>95%). Besides, the application of this image filtering system was shown to improve the diagnostic performance of established AI systems [121]. Automated assessment of images has also been explored in filtering OCT scans. The traditional index for assessing the quality of OCT scan is signal strength, which is limited in evaluating other factors such as off-centration, missing data, and artifacts. Ran et al. developed a 3D DL system for filtering out ungradable OCT volumes, with AUC of 0.954, sensitivity of 86.2%, and specificity of 92.6% using squeeze-and-excitation ResNeXt strategy [122]. Although DL algorithms for automated analysis of image quality are essential, cautions are warranted in setting the appropriate threshold, as images that are marked as ungradable by the DL system may result in unnecessary referral to the tertiary eye care centers.

Future Research and Challenges

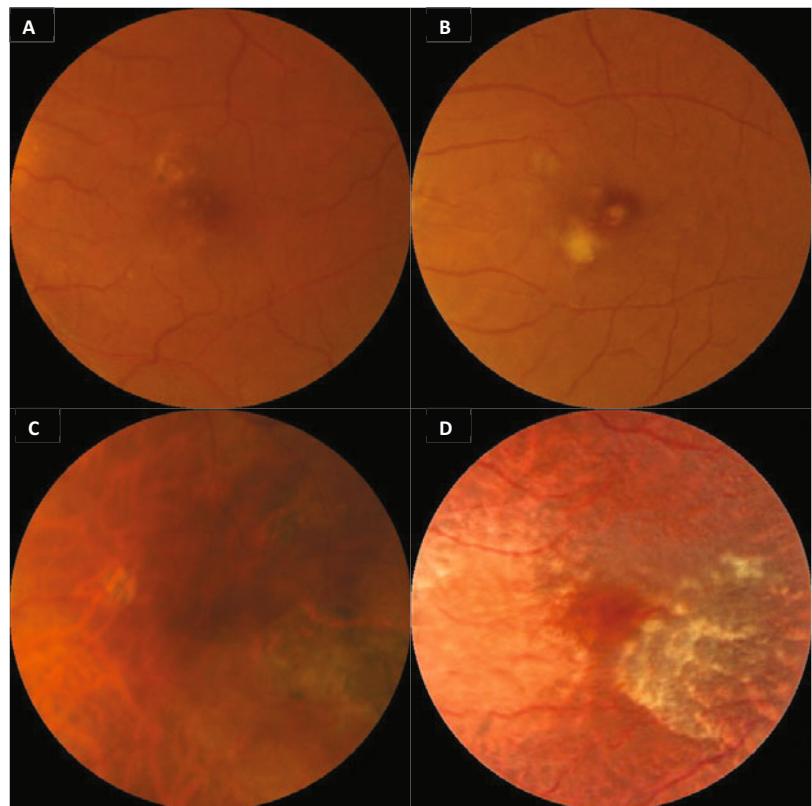
Novel Technical Approaches

At present, the majority of DL studies focus on segmentation and classification task. In terms of predictive DL algorithms, limited studies were conducted using the long-term clinical data to train the baseline images (fundus or OCTs) [123]. Although it may be ideal to develop more predictive algorithms to deliver preventive or personalized medicine, some diseases may simply do not have early signs that could be detected by the DL algorithms from the images. Even if they do, the limited longitudinal progression data is another contributory factor to suboptimal performance in many predictive algorithms. Another challenge related to

datasets is the availability of large training datasets, especially for rare diseases. To address this limitation, Burlina et al. applied low-shot DL methods to train the algorithm with relatively low number of images. Using 160 images, the DL model trained using low-shot methods and self-supervision outperformed the traditional DL system (AUC 0.747 vs 0.659) in differentiation of referable DR from non-referable DR. The superior performance remained when the training data was reduced further down to ten images. In addition, the low-shot methods might be beneficial to address AI bias due to imbalance of data or infeasibility of data partition [124].

Another innovative area in AI research is the application of generative adversarial network (GAN) to artificially synthesize ophthalmic images. GAN is an unsupervised DL machine based on two models: a generator and a discriminator, introduced by Goodfellow et al. [125]. The generative model learns to capture the data distribution, take random samples of noise, and generate plausible images from that distribution. The discriminative model estimates the probability that a sample comes from the data distribution rather than generator distribution, and therefore is tasked to discriminate between real and fake images. The performance of both networks simultaneously improves each other over time during the training process. GAN-synthesized images may be an alternative to address the limitations commonly encountered when developing a robust DL system, such as availability of large datasets, the privacy of patients, and personal data protection. Burlina et al. showed that the retinal specialists could not distinguish GAN-synthesized retinal images with AMD lesions from the real AMD images. Besides, the DL classifier trained on synthetic data achieved noninferior performance in detecting referable AMD (AUC 0.924) comparing to the one trained on real images (AUC 0.971) [126]. Figure 5 shows an example of GAN-synthesized normal fundus images and real images with AMD. Future research will be of great value to evaluate whether the similar algorithm could be generalized to other retinal conditions.

Fig. 5 Generative adversarial network (GAN) created images of AMD (B, D) compared to real images of AMD (A, C), with macular segmentation



Research Ethics and Artificial Images

Given the noninterventional and retrospective nature of AI research, institutional review boards in most countries do not require individual consent for such studies, particularly when the data being studied have been anonymized. For example, the NHS Constitution states that the physicians should harness the anonymized clinical information to support research and improve care for others [127]. The development of AI for clinical applications requires large amounts of training data, often an order of magnitude more than that collected even in large clinical trials. The application of AI in this context presents a number of considerations and challenges: 1) the datasets are often orders of magnitude larger than in previous retrospective case series, with attendant information security risks; 2) the datasets must often be shared outside the hospital in which they have been collected, either with academic or industry collaborators; and 3) whether or not ophthalmic images are considered truly

de-identified or the regulations around this may potentially change in the future.

Data Ownership and Sharing

In addition, the ownership and value of the data always raise a question regarding how the patients, institutions, and the public get the appropriate benefits from the use of their data to train AI systems? Although the relative value of clinical data is widely appreciated, it is far less clear how to determine its absolute value. For example, if an algorithm is trained on multiple datasets from different institutions or different countries, how can the relative contribution of each dataset be valued? In the coming years, the ophthalmic research community will learn more about the nuances of this – is a dataset of one million poor-quality retinal scans more valuable than a dataset of 100,000 scans of high quality? And what feature or disease distributions are optimal? A large dataset of healthy eyes is less likely to be

of value than a smaller one with a disease distribution more similar to that of the intended clinical application. These questions of value will become increasingly important as the use of innovative approaches such as federated learning become more widely adopted. Even when these factors have been elucidated, there are additional complexities, particularly when sharing data outside an academic setting. The governance around such decisions may be best undertaken by so-called “data trusts,” which allow the voices of multiple stakeholders to be considered and which can balance principles of reciprocity and proportionality [128, 129]. It appears increasingly likely that tiered access to data may be most appropriate, with larger companies being charged a license fee while smaller companies might prefer to offer a revenue share; conversely, academic researchers should ideally have free access to data for noncommercial applications.

Although the above considerations may be viewed negatively, as an additional burden on clinical researchers, in fact, there may be a step forward. If done correctly, the storage of data in cloud-based infrastructure may provide considerable additional safeguards against data loss, much better security, and greatly improved audit capabilities. With the introduction of European Union (EU) General Data Protection Regulations (GDPR), it will now be incumbent on clinical researchers to complete data protection impact assessments (DPIAs) before embarking on the use of innovative technologies, such as AI, in healthcare. This will provide an opportunity for researchers to incorporate good practice at an early stage in their studies and hopefully to avoid significant missteps at a much later stage. Taken together, extra care in these departments, combined with thoughtful engagement with patients will provide an opportunity to greatly improve patients trust and thus aid translation “from algorithm to application.”

Patients and Physicians Acceptance

To date, limited studies were available to assess the patients’ acceptance of the AI technologies in

medicine. In Tran et al., [130] merely half of the patients felt that the development of AI is an important opportunity, and 10% felt it could potentially dangerous. On the other hand, Keel et al. reported that 96% of participants reported that they were satisfied with the AI DR screening, with 78% preferred to have AI to manual screening [27]. One potential barrier to physicians’ acceptance to AI systems in ophthalmology is the “black box” functioning of these systems [131]. A number of technical methods (e.g., occlusion testing and integrated gradients) have been described to visualize where the DL “thinks” is abnormal. However, none of which has shown superiority to another [132, 133]. Regulatory bodies in the USA and Europe have also written about the potential importance of developing “explainable” AI systems [134–136]. Thus, to increase the adoption of AI within the clinical practice, the AI algorithms will need to be seamless, user-friendly, and unbiased to help making clinical decisions faster and more effectively [137, 138].

Education

The “tsunami” of the AI and digital health algorithms may warrant the educational department of medical schools and residency programs to revisit the teaching syllabus. To better equip the next generation of ophthalmologists with AI and digital health, it may be useful to include the ML and DL courses as part of the training, similar to how the statistics course is taught. Likewise, for the computer science or engineering department, it is worthwhile to explore the possibilities of having medical students or residents attachment to allow them to learn the basic courses on programming, coding, and statistics.

Guidelines

In April 2019, the US FDA adopted the definition by the International Medical Device Regulators Forum (IMDRF) to consider AI-based software as a medical device (SaMD) [139]. Applications of AI in ophthalmology described previously, for

example, were intended to be used for performing one or more ophthalmic purposes single-handedly without being part of a hardware medical device. The FDA also adopts quality systems and good ML practices for AI-based SaMD which include clinical evaluation guidance in three domains: valid clinical association between AI output and targeted condition, analytical validation (correct processing input data to generate accurate, reliable, and precise output data), and clinical validation (the accurate, reliable, and precise output data achieve the intended purposes in the targeted populations in clinical care).

In 2019, the World Health Organization (WHO) released a guideline to critically appraise digital health interventions such as AI, aiming to evaluate the benefits, harms, acceptability, feasibility, resource use, and equity considerations [140]. This not only helps prevent a proliferation of short-lived implementations of digital tools in the healthcare system, but also helps the health policymakers and relevant stakeholders to make informed investments into these large pools of AI algorithms in the market. The guideline also states that to promote digital health implementations, it is important to have the combination of health content, digital health interventions, and digital applications that can be supported by a robust information technology design and architecture.

To standardize the reporting of AI studies, an international effort has been made to develop AI extensions to the Consolidated Standards of Reporting Trials (CONSORT) and the Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) guidelines, which were described by Liu et al. and Rivera et al. with 14 and 15 additional AI-specific items, respectively [141, 142]. The CONSORT-AI and SPIRIT-AI extensions provide recommendations for reporting AI clinical trials, such as specifying the device used and indication for use, the applicable patient group, the type of data, and the limitation of the study.

In ophthalmology, AAO has also established an AI taskforce to gather the major AI key opinion leaders in ophthalmology, aiming to educate the current and next generation of

ophthalmologists on the AI-related research and clinical applications; [143] gather the experts' opinion to generate consensus guidelines on reporting metrics and regulations for different AI technologies; [17, 144–146] and explore the relatively uncertain safety and ethical considerations [147]. The establishment of guidelines, white papers, and recommendations will help streamline and increase quality of AI research and applications within the ophthalmology settings.

Conclusion

AI is an exciting, disruptive technology that will likely have a significant impact on medicine and healthcare. There has been remarkable progress in the use of AI in ophthalmology, and AI has been applied to many of the major eye diseases. However, there remains significant challenges for AI to show actual clinical translation in “real-world” settings to improve healthcare.

Acknowledgments We would like to acknowledge Ms. Valentina Bellemo and Xin Qi Lee, Singapore Eye Research Institute, for generating the tables and figures. Prof Hao Tian Lin for providing the figure on the articles. MFC is supported by National Institutes of Health grants R01EY19474 and P30EY10572, by National Science Foundation grant SCH-1622679, and by unrestricted departmental funding from Research to Prevent Blindness. Dr. Ting received grant support from National Medical Research Council (NMRC) Health Service Research Grant, Ministry of Health (MOH), Singapore (National Health Innovation Center, Innovation to Develop Grant (NHIC-I2-D-1409022); SingHealth Foundation Research Grant (SHF/FG648S/2015), and the Tanoto Foundation. The Singapore Diabetic Retinopathy Program (SiDRP) received funding from the MOH, Singapore (grants AIC/RPDD/SIDRP/SERI/FY2013/0018 & AIC/HPD/FY2016/0912). The Diabetes Study in Nephropathy and Other Microvascular Complications (DYNAMO) received funding from National Medical Research Council (NMRC) Large Collaborative Grant (LCG).

Financial Disclosure DT, CC, TYW are the patent holders of a deep learning system for retinal diseases. PK is the consultant for Google Deepmind, Roche and Novartis. MC is a Consultant for Novartis, and an equity owner in Inteloretina, LLC (Honolulu, HI).

References

1. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–44. <https://doi.org/10.1038/nature14539>.
2. Lee CS, Tyring AJ, Deruyter NP, Wu Y, Rokem A, Lee AY. Deep-learning based, automated segmentation of macular edema in optical coherence tomography. *Biomed Opt Express*. 2017;8(7):3440–8. <https://doi.org/10.1364/BOE.8.003440>.
3. Abramoff MD, Lou Y, Erginay A, Clarida W, Amelon R, Folk JC, et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Invest Ophthalmol Vis Sci*. 2016;57(13):5200–6. <https://doi.org/10.1167/iovs.16-19964>.
4. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402–10. <https://doi.org/10.1001/jama.2016.17216>.
5. Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multi-ethnic populations with diabetes. *JAMA*. 2017;318(22):2211–23. <https://doi.org/10.1001/jama.2017.18152>.
6. Gargyea R, Leng T. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*. 2017;124(7):962–9. <https://doi.org/10.1016/j.ophtha.2017.02.008>.
7. Li Z, He Y, Keel S, Meng W, Chang RT, He M. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology*. 2018;125(8):1199–206.
8. Burlina PM, Joshi N, Pekala M, Pacheco KD, Freund DE, Bressler NM. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmol*. 2017;135(11):1170–6.
9. Grassmann F, Mengelkamp J, Brandl C, Harsch S, Zimmermann ME, Linkohr B, et al. A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. *Ophthalmology*. 2018;125(9):1410–20. <https://doi.org/10.1016/j.ophtha.2018.02.037>.
10. Brown JM, Campbell JP, Beers A, Chang K, Ostmo S, Chan RVP, et al. Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. *JAMA Ophthalmol*. 2018;136(7):803–10. <https://doi.org/10.1001/jamaophthalmol.2018.1934>.
11. Varadarajan AV, Poplin R, Blumer K, Angermueller C, Ledsam J, Chopra R, et al. Deep learning for predicting refractive error from retinal fundus images. *Invest Ophthalmol Vis Sci*. 2018;59(7):2861–8. <https://doi.org/10.1167/iovs.18-23887>.
12. Milea D, Najjar RP, Zhubo J, Ting D, Vasseneix C, Xu X, et al. Artificial intelligence to detect papilledema from ocular fundus photographs. *N Engl J Med*. 2020;382(18):1687–95. <https://doi.org/10.1056/NEJMoa1917130>.
13. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*. 2018;2(3):158–64. <https://doi.org/10.1038/s41551-018-0195-0>.
14. Ting DSW, Cheung CY, Quang ND, Sabanayagam C, Lim G, Lim Z, et al. Deep learning in estimating prevalence and systemic risk factors for diabetic retinopathy: a multi-ethnic study. *NPJ Digit Med*. 2019;2:24.
15. Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*. 2018;172(5):1122–31.e9.
16. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. 2018;24(9):1342.
17. Ting DSW, Peng L, Varadarajan AV, Keane PA, Burlina PM, Chiang MF, et al. Deep learning in ophthalmology: the technical and clinical considerations. *Prog Retin Eye Res*. 2019;72:100759. <https://doi.org/10.1016/j.preteyes.2019.04.003>.
18. Ting DSW, Cheung GCM, Wong TY. Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review. *Clin Exp Ophthalmol*. 2016;44(4):260–77.
19. Cheung N, Mitchell P, Wong TY. Diabetic retinopathy. *Lancet*. 2010;376(9735):124–36. [https://doi.org/10.1016/S0140-6736\(09\)62124-3](https://doi.org/10.1016/S0140-6736(09)62124-3).
20. Yau JW, Rogers SL, Kawasaki R, Lamoureux EL, Kowalski JW, Bek T, et al. Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care*. 2012;35(3):556–64.
21. Wong TY, Bressler NM. Artificial intelligence with deep learning technology looks into diabetic retinopathy screening. *JAMA*. 2016;316(22):2366–7.
22. Scotland GS, McNamee P, Fleming AD, Goatman KA, Philip S, Prescott GJ, et al. Costs and consequences of automated algorithms versus manual grading for the detection of referable diabetic retinopathy. *Br J Ophthalmol*. 2010;94(6):712–9.
23. Abramoff MD, Lou Y, Erginay A, Clarida W, Amelon R, Folk JC, et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Invest Ophthalmol Vis Sci*. 2016;57(13):5200–6.
24. Li Z, Keel S, Liu C, He Y, Meng W, Scheetz J, et al. An automated grading system for detection of vision-threatening referable diabetic retinopathy on the basis of color fundus photographs. *Diabetes Care*. 2018;41(12):2509–16. <https://doi.org/10.2337/dc18-0147>.

25. Krause J, Gulshan V, Rahimy E, Karth P, Widner K, Corrado GS, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*. 2018;125(8):1264–72.
26. Abramoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med*. 2018;1(1):39.
27. Keel S, Lee PY, Scheetz J, Li Z, Kotowicz MA, MacIsaac RJ, et al. Feasibility and patient acceptability of a novel artificial intelligence-based screening model for diabetic retinopathy at endocrinology outpatient services: a pilot study. *Sci Rep*. 2018;8(1):4330. <https://doi.org/10.1038/s41598-018-22612-2>.
28. Bhuiyan A, Govindaiah A, Deobhakta A, Gupta M, Rosen R, Saleem S, et al. Development and validation of an automated diabetic retinopathy screening tool for primary care setting. *Diabetes Care*. 2020;43(10):e147–8. <https://doi.org/10.2337/dc19-2133>.
29. Varadarajan AV, Bavishi P, Ruamviboonsuk P, Chotcomwongse P, Venugopalan S, Narayanaswamy A, et al. Predicting optical coherence tomography-derived diabetic macular edema grades from fundus photographs using deep learning. *Nat Commun*. 2020;11(1):130. <https://doi.org/10.1038/s41467-019-13922-8>.
30. Ruamviboonsuk P, Krause J, Chotcomwongse P, Sayres R, Raman R, Widner K. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. *NPJ Digit Med*. 2019;2:Article Number 25.
31. Gulshan V, Rajan RP, Widner K, Wu D, Wubbels P, Rhodes T, et al. Performance of a deep-learning algorithm vs manual grading for detecting diabetic retinopathy in India. *JAMA Ophthalmol*. 2019;137:987.
32. Bellemo V, Lim ZW, Lim G, Nguyen QD, Xie Y, Yip MY, et al. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. *Lancet Digit Health*. 2019;1(1):e35–44.
33. Bora A, Balasubramanian S, Babenko B, Virmani S, Venugopalan S, Mitani A, et al. Predicting the risk of developing diabetic retinopathy using deep learning. *The Lancet Digital Health*. 2021;3(1):e10–e9. [https://doi.org/10.1016/S2589-7500\(20\)30250-8](https://doi.org/10.1016/S2589-7500(20)30250-8).
34. Jonas JB, Aung T, Bourne RR, Bron AM, Ritch R, Panda-Jonas S. *Glaucoma*. *Lancet*. 2017;390(10108):2183–93. [https://doi.org/10.1016/S0140-6736\(17\)31469-1](https://doi.org/10.1016/S0140-6736(17)31469-1).
35. Flaxman SR, Bourne RRA, Resnikoff S, Ackland P, Braithwaite T, Cincinelli MV, et al. Global causes of blindness and distance vision impairment 1990–2020: a systematic review and meta-analysis. *Lancet Glob Health*. 2017;5:e1221–34.
36. Tham Y-C, Li X, Wong TY, Quigley HA, Aung T, Cheng C-Y. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology*. 2014;121:2081–90.
37. Phene S, Dunn RC, Hammel N, Liu Y, Krause J, Kitade N, et al. Deep learning and glaucoma specialists: the relative importance of optic disc features to predict glaucoma referral in fundus photographs. *Ophthalmology*. 2019;126(12):1627–39. <https://doi.org/10.1016/j.ophtha.2019.07.024>.
38. Shibata N, Tanito M, Mitsuhashi K, Fujino Y, Matsuura M, Murata H, et al. Development of a deep residual learning algorithm to screen for glaucoma from fundus photography. *Sci Rep*. 2018;8(1):14665.
39. Masumoto H, Tabuchi H, Nakakura S, Ishitobi N, Miki M, Enno H. Deep-learning classifier with an ultrawide-field scanning laser ophthalmoscope detects glaucoma visual field severity. *J Glaucoma*. 2018;27(7):647–52.
40. Ko YC, Wey SY, Chen WT, Chang YF, Chen MJ, Chiou SH, et al. Deep learning assisted detection of glaucomatous optic neuropathy and potential designs for a generalizable model. *PLoS One*. 2020;15(5):e0233079. <https://doi.org/10.1371/journal.pone.0233079>.
41. Kim SJ, Cho KJ, Oh S. Development of machine learning models for diagnosis of glaucoma. *PLoS One*. 2017;12:e0177726.
42. Omodaka K, An G, Tsuda S, Shiga Y, Takada N, Kikawa T, et al. Classification of optic disc shape in glaucoma using machine learning based on quantified ocular parameters. *PLoS One*. 2017;12:e0190012.
43. Muhammad H, Fuchs TJ, De Cuir N, De Moraes CG, Blumberg DM, Liebmann JM, et al. Hybrid deep learning on single wide-field optical coherence tomography scans accurately classifies glaucoma suspects. *J Glaucoma*. 2017;26:1086.
44. Wong WL, Su X, Li X, Cheung CMG, Klein R, Cheng C-Y, et al. Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. *Lancet Glob Health*. 2014;2:e106–16.
45. Clemons TE, Milton RC, Klein R, Seddon JM. Risk factors for the incidence of advanced age-related macular degeneration in the Age-Related Eye Disease Study (AREDS) AREDS report no. 19. *Ophthalmology*. 2005;112:533–9.
46. Burlina PM, Joshi N, Pacheco KD, Freund DE, Kong J, Bressler NM. Use of deep learning for detailed severity characterization and estimation of 5-year risk among patients with age-related macular degeneration. *JAMA Ophthalmol*. 2018;136:1359–66.
47. Peng Y, Dharrsi S, Chen Q, Keenan TD, Agrón E, Wong WT, et al. DeepSeeNet: a deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs. *Ophthalmology*. 2019;126(4):565–75. <https://doi.org/10.1016/j.ophtha.2018.11.015>.
48. Keenan TD, Dharrsi S, Peng Y, Chen Q, Agrón E, Wong WT, et al. A deep learning approach for

- automated detection of geographic atrophy from color fundus photographs. *Ophthalmology*. 2019;126(11):1533–40. <https://doi.org/10.1016/j.ophtha.2019.06.005>.
49. Liefers B, Colijn JM, González-Gonzalo C, Verzijden T, Wang JJ, Joachim N, et al. A deep learning model for segmentation of geographic atrophy to study its Long-term natural history. *Ophthalmology*. 2020. <https://doi.org/10.1016/j.ophtha.2020.02.009>.
50. Bhuiyan A, Wong TY, Ting DSW, Govindaiah A, Souied EH, Smith RT. Artificial intelligence to stratify severity of age-related macular degeneration (AMD) and predict risk of progression to late AMD. *Transl Vis Sci Technol*. 2020;9(2):25. <https://doi.org/10.1167/tvst.9.2.25>.
51. Lee CS, Baughman DM, Lee AY. Deep learning is effective for classifying normal versus age-related macular degeneration OCT images. *Ophthalmol Retina*. 2017;1(4):322–7.
52. Treder M, Lauermann JL, Eter N. Automated detection of exudative age-related macular degeneration in spectral domain optical coherence tomography using deep learning. *Graefes Arch Clin Exp Ophthalmol*. 2018;256(2):259–65. <https://doi.org/10.1007/s00417-017-3850-3>.
53. Pascolini D, Mariotti SP. Global estimates of visual impairment: 2010. *Br J Ophthalmol*. 2012;96(5):614–8. <https://doi.org/10.1136/bjophthalmol-2011-300539>.
54. Fleck BW, Dangata Y. Causes of visual handicap in the Royal Blind School, Edinburgh, 1991–2. *Br J Ophthalmol*. 1994;78(5):421.
55. Redd TK, Campbell JP, Brown JM, Kim SJ, Ostmo S, Chan RVP, et al. Evaluation of a deep learning image assessment system for detecting severe retinopathy of prematurity. *Br J Ophthalmol*. 2019;103(5):580–4.
56. Greenwald MF, Danford ID, Shahrawat M, Ostmo S, Brown J, Kalpathy-Cramer J, et al. Evaluation of artificial intelligence-based telemedicine screening for retinopathy of prematurity. *J AAPOS*. 2020;24(3):160–2. <https://doi.org/10.1016/j.jaapos.2020.01.014>.
57. Tan Z, Simkin S, Lai C, Dai S. Deep learning algorithm for automated diagnosis of retinopathy of prematurity plus disease. *Transl Vis Sci Technol*. 2019;8(6):23. <https://doi.org/10.1167/tvst.8.6.23>.
58. Bruce BB, Lamirel C, Wright DW, Ward A, Heilpern KL, Bioussé V, et al. Nonmydriatic ocular fundus photography in the emergency department. *N Engl J Med*. 2011;364(4):387–9. <https://doi.org/10.1056/NEJMc1009733>.
59. Mackay DD, Garza PS, Bruce BB, Newman NJ, Bioussé V. The demise of direct ophthalmoscopy: a modern clinical challenge. *Neurol Clin Pract*. 2015;5(2):150–7. <https://doi.org/10.1212/cpj.0000000000000115>.
60. Irani NK, Bidot S, Peragallo JH, Esper GJ, Newman NJ, Bioussé V. Feasibility of a nonmydriatic ocular fundus camera in an outpatient neurology clinic. *Neurologist*. 2020;25(2):19–23. <https://doi.org/10.1097/nrl.0000000000000259>.
61. Ivan Y, Ramgopal S, Cardenas-Villa M, Winger DG, Wang L, Vitale MA, et al. Feasibility of the digital retinography system camera in the pediatric emergency department. *Pediatr Emerg Care*. 2018;34(7):488–91. <https://doi.org/10.1097/pec.0000000000001203>.
62. Bioussé V, Newman NJ, Najjar RP, Vasseneix C, Xu X, Ting DS, et al. Optic disc classification by deep learning versus expert neuro-ophthalmologists. *Ann Neurol*. 2020. <https://doi.org/10.1002/ana.25839>.
63. Ahn JM, Kim S, Ahn KS, Cho SH, Kim US. Accuracy of machine learning for differentiation between optic neuropathies and pseudopapilledema. *BMC Ophthalmol*. 2019;19(1):178. <https://doi.org/10.1186/s12886-019-1184-0>.
64. Roth GA, Johnson C, Abajobir A, Abd-Allah F, Abera SF, Abyu G, et al. Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. *J Am Coll Cardiol*. 2017;70(1):1–25. <https://doi.org/10.1016/j.jacc.2017.04.052>.
65. Lanza GA, Crea F. Primary coronary microvascular dysfunction: clinical presentation, pathophysiology, and management. *Circulation*. 2010;121(21):2317–25. <https://doi.org/10.1161/CIRCULATIONAHA.109.900191>.
66. Guterman DD, Chabowski DS, Kadlec AO, Durand MJ, Freed JK, Ait-Aissa K, et al. The human microcirculation: regulation of flow and beyond. *Circ Res*. 2016;118(1):157–72. <https://doi.org/10.1161/CIRCRESAHA.115.305364>.
67. Strain WD, Paldanius PM. Diabetes, cardiovascular disease and the microcirculation. *Cardiovasc Diabetol*. 2018;17(1):57. <https://doi.org/10.1186/s12933-018-0703-2>.
68. Keith NM, Wagener HP, Barker NW. Some different types of essential hypertension: their course and prognosis. *Am J Med Sci*. 1939;197(3):332–43.
69. Wong TY, Mitchell P. Hypertensive retinopathy. *N Engl J Med*. 2004;351(22):2310–7.
70. Williams B, Poulter NR, Brown MJ, Davis M, McInnes GT, Potter JF, et al. British hypertension society guidelines for hypertension management 2004 (BHS-IV): summary. *BMJ*. 2004;328(7440):634–40. <https://doi.org/10.1136/bmj.328.7440.634>.
71. Chobanian AV, Bakris GL, Black HR, Cushman WC, Green LA, Izzo JL Jr, et al. The seventh report of the joint National Committee on prevention, detection, evaluation, and treatment of high blood pressure: the JNC 7 report. *JAMA*. 2003;289(19):2560–72. <https://doi.org/10.1001/jama.289.19.2560>.
72. Mansia G, De Backer G, Dominiczak A, Cifkova R, Fagard R, Germano G, et al. 2007 ESH-ESC guidelines for the management of arterial hypertension: the task force for the management of arterial hypertension of the European Society of Hypertension (ESH) and

- of the European Society of Cardiology (ESC). *Blood Press.* 2007;16(3):135–232. <https://doi.org/10.1080/08037050701461084>.
73. Hypertension in adults: diagnosis and management NICE Clinical guideline [CG127]. 2011. <https://www.nice.org.uk/guidance/cg127/chapter/1-Guidance#assessing-cardiovascular-risk-and-target-organ-damage>. Accessed Feb 2019.
74. Ting DSW, Wong TY. Eyeing cardiovascular risk factors. *Nat Biomed Eng.* 2018;2(3):140–1. <https://doi.org/10.1038/s41551-018-0210-5>.
75. Rim TH, Lee G, Kim Y, Tham Y-C, Lee CJ, Baik SJ, et al. Prediction of systemic biomarkers from retinal photographs: development and validation of deep-learning algorithms. *Lancet Digit Health.* 2020;2(10):e526–36. [https://doi.org/10.1016/S2589-7500\(20\)30216-8](https://doi.org/10.1016/S2589-7500(20)30216-8).
76. Cheung CY, Xu D, Cheng CY, Sabanayagam C, Tham YC, Yu M, et al. A deep-learning system for the assessment of cardiovascular disease risk via the measurement of retinal-vessel calibre. *Nat Biomed Eng.* 2020. <https://doi.org/10.1038/s41551-020-00626-4>.
77. Chang J, Ko A, Park SM, Choi S, Kim K, Kim SM, et al. Association of cardiovascular mortality and deep learning-funduscopic atherosclerosis score derived from retinal fundus images. *Am J Ophthalmol.* 2020;217:121–30. <https://doi.org/10.1016/j.ajo.2020.03.027>.
78. Son J, Shin JY, Chun EJ, Jung K-H, Park KH, Park SJ. Predicting high coronary artery calcium score from retinal fundus images with deep learning algorithms. *Transl Vis Sci Technol.* 2020;9(2):28. <https://doi.org/10.1167/tvst.9.2.28>.
79. Dai G, He W, Xu L, Pazó EE, Lin T, Liu S, et al. Exploring the effect of hypertension on retinal microvasculature using deep learning on East Asian population. *PLoS One.* 2020;15(3):e0230111. <https://doi.org/10.1371/journal.pone.0230111>.
80. Zhang L, Yuan M, An Z, Zhao X, Wu H, Li H, et al. Prediction of hypertension, hyperglycemia and dyslipidemia from retinal fundus photographs via deep learning: a cross-sectional study of chronic diseases in central China. *PLoS One.* 2020;15(5):e0233166. <https://doi.org/10.1371/journal.pone.0233166>.
81. Mitani A, Huang A, Venugopalan S, Corrado GS, Peng L, Webster DR, et al. Detection of anaemia from retinal fundus images via deep learning. *Nat Biomed Eng.* 2020;4(1):18–27. <https://doi.org/10.1038/s41551-019-0487-z>.
82. Sabanayagam C, Xu D, Ting DSW, Nusinovici S, Banu R, Hamzah H, et al. A deep learning algorithm to detect chronic kidney disease from retinal photographs in community-based populations. *Lancet Digit Health.* 2020;2(6):e295–302. [https://doi.org/10.1016/S2589-7500\(20\)30063-7](https://doi.org/10.1016/S2589-7500(20)30063-7).
83. Fujimoto J, Swanson E. The development, commercialization, and impact of optical coherence tomography. *Invest Ophthalmol Vis Sci.* 2016;57(9):OCT1–OCT13.
84. Keane PA, Patel PJ, Liakopoulos S, Heussen FM, Sadda SR, Tufail A. Evaluation of age-related macular degeneration with optical coherence tomography. *Surv Ophthalmol.* 2012;57(5):389–414.
85. Keane PA, Sadda SR. Optical coherence tomography in the diagnosis and management of diabetic retinopathy. *Int Ophthalmol Clin.* 2009;49(2):61–74.
86. Windsor MA, Sun SJ, Frick KD, Swanson EA, Rosenfeld PJ, Huang D. Estimating public and patient savings from basic research – a study of optical coherence tomography in managing antiangiogenic therapy. *Am J Ophthalmol.* 2018;185:115–22.
87. Waldstein SM, Simader C, Staurenghi G, Chong NV, Mitchell P, Jaffe GJ, et al. Morphology and visual acuity in aflibercept and ranibizumab therapy for neovascular age-related macular degeneration in the VIEW trials. *Ophthalmology.* 2016;123(7):1521–9.
88. Schlegl T, Waldstein SM, Bogunovic H, Endstraßer F, Sadeghipour A, Philip A-M, et al. Fully automated detection and quantification of macular fluid in OCT using deep learning. *Ophthalmology.* 2018;125(4):549–58.
89. Keenan TDL, Clemons TE, Domalpally A, Elman MJ, Havilio M, Agrón E, et al. Retinal specialist versus artificial intelligence detection of retinal fluid from OCT: age-related eye disease study 2: 10-year follow-on study. *Ophthalmology.* 2020. <https://doi.org/10.1016/j.ophtha.2020.06.038>.
90. Saha S, Nassisi M, Wang M, Lindenberg S, Kanagasingam Y, Sadda S, et al. Automated detection and classification of early AMD biomarkers using deep learning. *Sci Rep.* 2019;9(1):10990. <https://doi.org/10.1038/s41598-019-47390-3>.
91. Rim TH, Lee AY, Ting DS, Teo K, Betzler BK, Teo ZL, et al. Detection of features associated with neovascular age-related macular degeneration in ethnically distinct data sets by an optical coherence tomography: trained deep learning algorithm. *Br J Ophthalmol.* 2020. <https://doi.org/10.1136/bjophthalmol-2020-316984>.
92. Schmidt-Erfurth U, Vogl WD, Jampol LM, Bogunović H. Application of automated quantification of fluid volumes to anti-VEGF therapy of neovascular age-related macular degeneration. *Ophthalmology.* 2020;127(9):1211–9. <https://doi.org/10.1016/j.ophtha.2020.03.010>.
93. Schmidt-Erfurth U, Bogunovic H, Sadeghipour A, Schlegl T, Langs G, Gerendas BS, et al. Machine learning to analyze the prognostic value of current imaging biomarkers in neovascular age-related macular degeneration. *Ophthalmol Retina.* 2018;2(1):24–30.
94. Yim J, Chopra R, Spitz T, Winkens J, Obika A, Kelly C, et al. Predicting conversion to wet age-related macular degeneration using deep learning. *Nat Med.* 2020;26(6):892–9. <https://doi.org/10.1038/s41591-020-0867-7>.

95. Vaghefi E, Hill S, Kersten HM, Squirrell D. Multimodal retinal image analysis via deep learning for the diagnosis of intermediate dry age-related macular degeneration: a feasibility study. *J Ophthalmol*. 2020;2020:7493419. <https://doi.org/10.1155/2020/7493419>.
96. Klein BE, Johnson CA, Meuer SM, Lee K, Wahle A, Lee KE, et al. Nerve fiber layer thickness and characteristics associated with glaucoma in community living older adults: prelude to a screening trial? *Ophthalmic Epidemiol*. 2017;24(2):104–10. <https://doi.org/10.1080/09286586.2016.1258082>.
97. Liu MM, Cho C, Jefferys JL, Quigley HA, Scott AW. Use of optical coherence tomography by non-expert personnel as a screening approach for glaucoma. *J Glaucoma*. 2018;27(1):64–70. <https://doi.org/10.1097/IJG.0000000000000822>.
98. Medeiros FA, Jammal AA, Thompson AC. From machine to machine: an OCT-trained deep learning algorithm for objective quantification of glaucomatous damage in fundus photographs. *Ophthalmology*. 2019;126(4):513–21. <https://doi.org/10.1016/j.ophtha.2018.12.033>.
99. Thompson AC, Jammal AA, Medeiros FA. A deep learning algorithm to quantify neuroretinal rim loss from optic disc photographs. *Am J Ophthalmol*. 2019;201:9–18. <https://doi.org/10.1016/j.ajo.2019.01.011>.
100. Medeiros FA, Jammal AA, Mariottoni EB. Detection of progressive glaucomatous optic nerve damage on fundus photographs with deep learning. *Ophthalmology*. 2020. <https://doi.org/10.1016/j.ophtha.2020.07.045>.
101. Asaoka R, Murata H, Hirasawa K, Fujino Y, Matsura M, Miki A, et al. Using deep learning and transfer learning to accurately diagnose early-onset glaucoma from macular optical coherence tomography images. *Am J Ophthalmol*. 2019;198:136–45. <https://doi.org/10.1016/j.ajo.2018.10.007>.
102. Kim KE, Kim JM, Song JE, Kee C, Han JC, Hyun SH. Development and validation of a deep learning system for diagnosing glaucoma using optical coherence tomography. *J Clin Med*. 2020;9(7):2167. <https://doi.org/10.3390/jcm9072167>.
103. Ran AR, Cheung CY, Wang X, Chen H, Luo L-Y, Chan PP, et al. Detection of glaucomatous optic neuropathy with spectral-domain optical coherence tomography: a retrospective training and validation deep-learning analysis. *Lancet Digit Health*. 2019;1(4):e172–82. [https://doi.org/10.1016/S2589-7500\(19\)30085-8](https://doi.org/10.1016/S2589-7500(19)30085-8).
104. Medeiros FA. Deep learning in glaucoma: progress, but still lots to do. *Lancet Digit Health*. 2019;1(4): e151–2. [https://doi.org/10.1016/S2589-7500\(19\)30087-1](https://doi.org/10.1016/S2589-7500(19)30087-1).
105. Ang M, Baskaran M, Werkmeister RM, Chua J, Schmidl D, Aranha Dos Santos V, et al. Anterior segment optical coherence tomography. *Prog Retin Eye Res*. 2018;66:132–56. <https://doi.org/10.1016/j.preteyeres.2018.04.002>.
106. Lai I, Mak H, Lai G, Yu M, Lam DS, Leung CK. Anterior chamber angle imaging with swept-source optical coherence tomography: measuring peripheral anterior synechia in glaucoma. *Ophthalmology*. 2013;120(6):1144–9. <https://doi.org/10.1016/j.ophtha.2012.12.006>.
107. Xu BY, Chiang M, Chaudhary S, Kulkarni S, Pardeshi AA, Varma R. Deep learning classifiers for automated detection of gonioscopic angle closure based on anterior segment OCT images. *Am J Ophthalmol*. 2019;208:273–80.
108. He M, Jiang Y, Huang S, Chang DS, Munoz B, Aung T, et al. Laser peripheral iridotomy for the prevention of angle closure: a single-centre, randomised controlled trial. *Lancet*. 2019;393(10181):1609–18. [https://doi.org/10.1016/S0140-6736\(18\)32607-2](https://doi.org/10.1016/S0140-6736(18)32607-2).
109. Li F, Wang Z, Qu G, Song D, Yuan Y, Xu Y, et al. Automatic differentiation of glaucoma visual field from non-glaucoma visual field using deep convolutional neural network. *BMC Med Imaging*. 2018;18(1):35. <https://doi.org/10.1186/s12880-018-0273-5>.
110. Yousefi S, Goldbaum MH, Balasubramanian M, Medeiros FA, Zangwill LM, Liebmann JM, et al. Learning from data: recognizing glaucomatous defect patterns and detecting progression from visual field measurements. *IEEE Trans Biomed Eng*. 2014;61(7):2112–24. <https://doi.org/10.1109/TBME.2014.2314714>.
111. Garcia G-GP, Nitta K, Lavieri MS, Andrews C, Liu X, Lobaza E, et al. Using Kalman filtering to forecast disease trajectory for patients with normal tension glaucoma. *Am J Ophthalmol*. 2019;199:111–9.
112. Wen JC, Lee CS, Keane PA, Xiao S, Rokem AS, Chen PP, et al. Forecasting future Humphrey visual fields using deep learning. *PLoS One*. 2019;14(4):e0214875. <https://doi.org/10.1371/journal.pone.0214875>.
113. Mariottoni EB, Datta S, Dov D, Jammal AA, Berchuck SI, Tavares IM, et al. Artificial intelligence mapping of structure to function in glaucoma. *Transl Vis Sci Technol*. 2020;9(2):19. <https://doi.org/10.1167/tvst.9.2.19>.
114. Wang M, Shen LQ, Pasquale LR, Wang H, Li D, Choi EY, et al. An artificial intelligence approach to assess spatial patterns of retinal nerve fiber layer thickness maps in glaucoma. *Transl Vis Sci Technol*. 2020;9(9):41. <https://doi.org/10.1167/tvst.9.9.41>.
115. Park K, Kim J, Lee J. A deep learning approach to predict visual field using optical coherence tomography. *PLoS One*. 2020;15(7):e0234902. <https://doi.org/10.1371/journal.pone.0234902>.
116. Long E, Liu Z, Xiang Y, Xu A, Huang J, Huang X, et al. Discrimination of the behavioural dynamics of visually impaired infants via deep learning. *Nat Biomed Eng*. 2019;3:860–9.
117. Jha AK, Des Roches CM, Campbell EG, Donelan K, Rao SR, Ferris TG, et al. Use of electronic health

- records in U.S. hospitals. *N Engl J Med.* 2009;360(16):1628–38. <https://doi.org/10.1056/NEJMsa0900592>.
118. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med.* 2018;1:18. <https://doi.org/10.1038/s41746-018-0029-1>.
119. Baxter SL, Marks C, Kuo T-T, Ohno-Machado L, Weinreb RN. Machine learning-based predictive modeling of surgical intervention in glaucoma using systemic data from electronic health records. *Am J Ophthalmol.* 2019;208:30–40.
120. Coyner AS, Swan R, Campbell JP, Ostmo S, Brown JM, Kalpathy-Cramer J, et al. Automated fundus image quality assessment in retinopathy of prematurity using deep convolutional neural networks. *Ophthalmol Retina.* 2019;3(5):444–50. <https://doi.org/10.1016/j.oret.2019.01.015>.
121. Li Z, Guo C, Nie D, Lin D, Zhu Y, Chen C, et al. Deep learning from “passive feeding” to “selective eating” of real-world data. *NPJ Digit Med.* 2020;3:143. <https://doi.org/10.1038/s41746-020-00350-y>.
122. Ran AR, Shi J, Ngai AK, Chan WY, Chan PP, Young AL, et al. Artificial intelligence deep learning algorithm for discriminating ungradable optical coherence tomography three-dimensional volumetric optic disc scans. *Neurophotonics.* 2019;6(4):041110. <https://doi.org/10.1117/1.NPh.6.4.041110>.
123. Schmidt-Erfurth U, Waldstein SM, Klimscha S, Sadeghipour A, Hu X, Gerendas BS, et al. Prediction of individual disease conversion in early AMD using artificial intelligence. *Invest Ophthalmol Vis Sci.* 2018;59(8):3199–208. <https://doi.org/10.1167/iovs.18-24106>.
124. Burlina P, Paul W, Mathew P, Joshi N, Pacheco KD, Bressler NM. Low-shot deep learning of diabetic retinopathy with potential applications to address artificial intelligence bias in retinal diagnostics and rare ophthalmic diseases. *JAMA Ophthalmol.* 2020;138(10):1070–7. <https://doi.org/10.1001/jamaophthalmol.2020.3269>.
125. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. editors. Generative adversarial nets. Advances in neural information processing systems; 2014.
126. Burlina PM, Joshi N, Pacheco KD, Liu TYA, Bressler NM. Assessment of deep generative models for high-resolution synthetic retinal image generation of age-related macular degeneration. *JAMA Ophthalmol.* 2019;137(3):258–64. <https://doi.org/10.1001/jamaophthalmol.2018.6156>.
127. England N. The NHS constitution. The NHS belongs to us all. London: NHS England; 2015.
128. Data Trusts: Lessons from Three Pilots (Report). 2019. <https://theodi.org/article/odi-data-trusts-report/>. Accessed 17 Oct 2019.
129. Price WN, Cohen IG. Privacy in the age of medical big data. *Nat Med.* 2019;25(1):37–43.
130. Tran VT, Riveros C, Ravaud P. Patients’ views of wearable devices and AI in healthcare: findings from the ComPaRe e-cohort. *NPJ Digit Med.* 2019;2:53. <https://doi.org/10.1038/s41746-019-0132-y>.
131. Castelvecchi D. Can we open the black box of AI? *Nature.* 2016;538(7623):20–3. <https://doi.org/10.1038/538020a>.
132. Hohman F, Kahng M, Pienta R, Chau DH. Visual analytics in deep learning: an interrogative survey for the next frontiers. 2018. <https://arxiv.org/pdf/1801.06889.pdf>. Accessed 10 Nov 2019.
133. Shortliffe EH, Sepulveda MJ. Clinical decision support in the era of artificial intelligence. *JAMA.* 2018;320(21):2199–200. <https://doi.org/10.1001/jama.2018.17163>.
134. Younis N, Broadbent DM, Vora JP, Harding SP. Incidence of sight-threatening retinopathy in patients with type 2 diabetes in the Liverpool Diabetic Eye Study: a cohort study. *Lancet.* 2003;361(9353):195–200.
135. Klein R, Klein BE, Moss SE, Cruickshanks KJ. The Wisconsin Epidemiologic Study of Diabetic Retinopathy. XV. The long-term incidence of macular edema. *Ophthalmology.* 1995;102(1):7–16.
136. Maguire A, Chan A, Cusumano J, Hing S, Craig M, Silink M, et al. The case for biennial retinopathy screening in children and adolescents. *Diabetes Care.* 2005;28(3):509–13.
137. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data.* 2016;3:160035. <https://doi.org/10.1038/sdata.2016.35>.
138. Celi LA, Fine B, Stone DJ. An awakening in medicine: the partnership of humanity and intelligent machines. *Lancet Digit Health.* 2019;1(6):e255–7.
139. Administration UFAD. Proposed regulatory framework for modifications for artificial intelligence/machine learning – based software as a medical device (SaMD). 2019. <https://www.fda.gov/media/122535/download>. Accessed 17 Aug 2019.
140. Organization WH. WHO guideline: recommendations on digital interventions for health system strengthening: web supplement 2: summary of findings and GRADE tables. World Health Organization; 2019.
141. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, Chan A-W, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med.* 2020;26(9):1364–74. <https://doi.org/10.1038/s41591-020-1034-x>.
142. Rivera SC, Liu X, Chan A-W, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *BMJ.* 2020;370:m3210. <https://doi.org/10.1136/bmj.m3210>.
143. Ting DSW, Lee AY, Wong TY. An ophthalmologist’s guide to deciphering studies in artificial intelligence. *Ophthalmology.* 2019;126(11):1475–9. <https://doi.org/10.1016/j.ophtha.2019.09.014>.

144. Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, et al. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol*. 2019;103(2):167–75.
145. Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, Waldstein SM, Bogunovic H. Artificial intelligence in retina. *Prog Retin Eye Res*. 2018;67:1–29. <https://doi.org/10.1016/j.preteyeres.2018.07.004>.
146. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet*. 2019;393(10181):1577–9. [https://doi.org/10.1016/S0140-6736\(19\)30037-6](https://doi.org/10.1016/S0140-6736(19)30037-6).
147. Char DS, Shah NH, Magnus D. Implementing machine learning in health care – addressing ethical challenges. *N Engl J Med*. 2018;378(11):981–3. <https://doi.org/10.1056/NEJMp1714229>.



Artificial Intelligence in Ophthalmology

111

Technical and Clinical Uses and Clinical Practice Challenges

Leonardo Seidi Shigueoka, Alessandro Adad Jammal,
Felipe Andrade Medeiros, and Vital Paulino Costa

Contents

Introduction	1553
Clinical Application of AI in Ophthalmology	1554
Diabetic Retinopathy (DR)	1555
Glaucoma	1556
Age-Related Macular Degeneration (AMD)	1561
Conclusions	1561
References	1564

Abstract

Ophthalmology presents as an exciting field in medicine for artificial intelligence (AI) systems, with its numerous digital imaging techniques – only surpassed, perhaps, by radiology – and the increasing prevalence of eye diseases that cause preventable vision loss that accompanies the global increase in life expectancy. Machine learning and deep learning can be applied for screening, detection, identification of progression, and assessment of the response to treatment of the main eye diseases. Recent studies demonstrated that these systems show good

performance in the detection of diabetic retinopathy, glaucoma, age-related macular degeneration, retinopathy of prematurity, refractive errors, and in the identification of risk factors for systemic diseases using eye fundus photos. This chapter describes the technical and clinical uses of AI in ophthalmology and discusses the challenges for its incorporation into clinical practice.

Keywords

Artificial intelligence · Machine learning · Deep learning · Big data

Introduction

To date, the application of data science to ophthalmology has been less apparent than in other areas. However, large clinical databases obtained from electronic medical records and digital

L. S. Shigueoka · A. A. Jammal · V. P. Costa (✉)
University of Campinas, São Paulo, Brazil
e-mail: alessandro.jammal@duke.edu

F. A. Medeiros
Department of Ophthalmology, Duke Eye Center, Durham,
NC, USA
e-mail: felipe.medeiros@duke.edu

images have provided opportunities for the detection of several ocular diseases through artificial intelligence (AI).

Diabetic retinopathy (DR), age-macular degeneration (AMD), glaucoma, and cataract represent the main causes of blindness worldwide, with alarming projections of increased prevalence with aging [1]. Over the past decade, AI has become a popular topic, since substantial burden of visual impairment and blindness from major eye diseases have not been adequately tackled using traditional models of eye care [2, 3].

New algorithms, systems, and software have been developed to perform diagnostic and grading tasks from imaging with operations analogous to human's learning and decision-making. In addition, accurate algorithms for pattern recognition may help clinicians by reducing diagnostic and therapeutic errors with the interpretation of large quantity of complex data in time-consuming tasks.

AI has three main applications in imaging: (1) Classification, where an image will be classified into different categories, for example, presence or absence of disease, or stage of the disease. This function can be used for automated diagnosis, screening, or staging. (2) Segmentation, to detect and outline anatomical structures or injuries, in order to measure shape or volume. This can be used for automated quantification of biomarkers in an image. (3) Prediction, to predict future results or predict the value of another measure. For example, predicting visual acuity, age, or blood pressure from an image. This function can also be used to estimate the prognosis of diseases or to promote correlation between structure and function.

The practical utility of an algorithm is to aid clinicians on decision-making and to educate or advise patients. There are two main modalities of AI's clinical utility: (1) computer-aided diagnosis allows clinicians to use the computer's output data as a "second opinion," which will help on decision-making; and (2) automated computer diagnosis, which suggests a diagnostic classification by the algorithm output, without the input of a physician.

Clinical Application of AI in Ophthalmology

The outpatient nature of the practice and the use of various imaging modalities makes ophthalmology a promising field to the implementation of AI for screening, diagnostic staging, and therapeutic guidance.

In 2050, the world population over 60 is estimated to reach 2 billion. The increase in longevity will lead to an increase in the number of people with visual impairment and blindness caused by diabetic retinopathy, glaucoma, and AMD [4]. Other eye diseases that require early detection and are important causes of childhood blindness include retinopathy of prematurity, refractive errors, and amblyopia [5]. AI algorithms can be used as alternative screening tools for all the abovementioned eye diseases.

The most common application of AI methods to the retina includes the detection of disease-related features from color fundus photographs. Fundus photography is an important screening tool for eye diseases, due to the low cost, portability, and wide availability of fundus cameras, allowing *in vivo* evaluation of the microvasculature and neural tissue. An important first step in evaluating fundus photographs is to identify whether the image orientation is adequate so that the automated system can analyze the condition of the retina. The landmarks of the retina commonly used for this task are the large retinal vessels, the optic disc, and the fovea, which can be easily recognized (Fig. 1) [6]. Using fundus photographs, AI has shown good accuracy in predicting age, gender, blood pressure, and smoking habit in an individual. The same algorithm was also able to indicate whether an individual had a higher risk for unstable angina, heart attack, stroke, or death due to cardiovascular causes [7].

With the increasing success of neural networks, there is a need to explain its decision base. The interpretability of the neural network generally is given by the identification of which part of an image is responsible for the activation of the network. This is typically represented in the form of a heat map, which indicates the location of

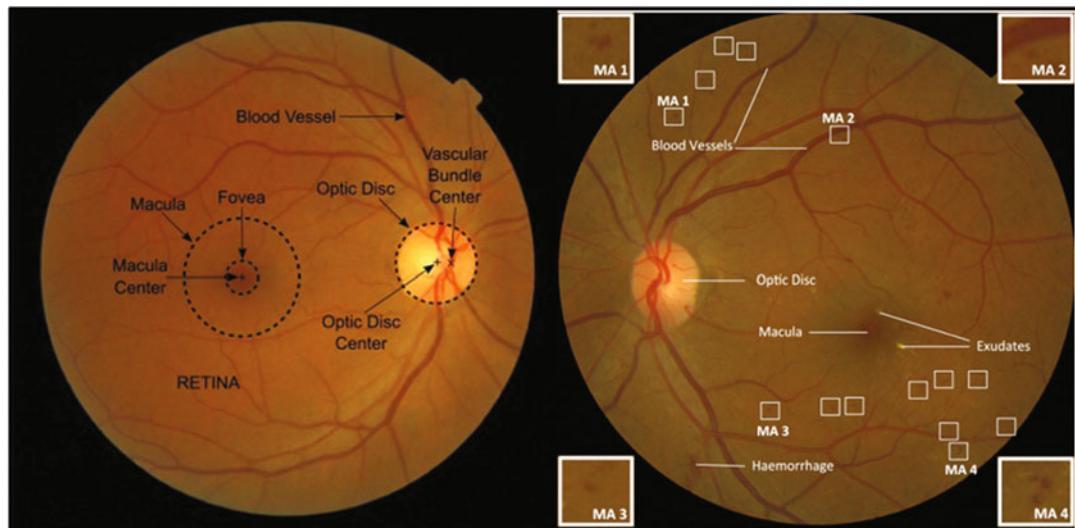


Fig. 1 Left: Anatomical landmarks used for orientation in a fundus photograph: fovea, macula, blood vessels, optic nerve head, and optic nerve head center. Right: Detection

the image where the features modify the network's predictions. The most common and simplest approach is to perform the so-called occlusion test (Fig. 2) [8]. To identify the areas of the image that contribute most to the decision of the neural network, a "white box" is placed in all areas of the image and the respective probability of output class is recorded. The biggest drop in the likelihood of categorization will represent the region of greatest importance.

The following sections will discuss the use of AI for the screening, diagnosis, and prognosis of diabetic retinopathy, glaucoma, and age-related macular degeneration.

Diabetic Retinopathy (DR)

Diabetic retinopathy is a specific microvascular complication of diabetes and remains one of the leading causes of preventable blindness in the world. It is identified in one-third of people with diabetes and it is associated with increased risk of life-threatening systemic vascular events, such as stroke, coronary heart disease, and heart failure [9]. Although diabetes affects the eye in many ways (e.g., increased risk of cataract), diabetic

of abnormal structures such as microaneurysms and hemorrhages; the white squares show examples of microaneurysms detected by the algorithm [6]

retinopathy is the most common and serious ocular complication [10].

Ophthalmologists typically diagnose and stage DR by the direct assessment of the retina during a fundus examination or by color fundus photographs. Given the large number of diabetes patients globally and to address the shortfalls of current diagnostic workflows, automated solutions could ease the burden of expensive and time-consuming screening by specialized doctors. A clinical trial demonstrated a sensitivity of 87.2% and a specificity of 90.7% in the detection of DR by AI, with a classification accuracy of 96.1%, using an external reading center as a reference standard. It is interesting to note that this reading center used wide-angle retinal examination and optical coherence tomography (OCT) to detect the presence of macular edema, whereas the algorithms were developed with the use of inexpensive posterior pole retinography. Due to differences in reference standards, it is difficult to compare the performance of AI algorithms reported in several studies [11]. To solve this challenge, an algorithm was tested on 11 independent datasets with a single reference standard. This algorithm showed a sensitivity of 90.5% and a specificity of 91.6% in the primary dataset

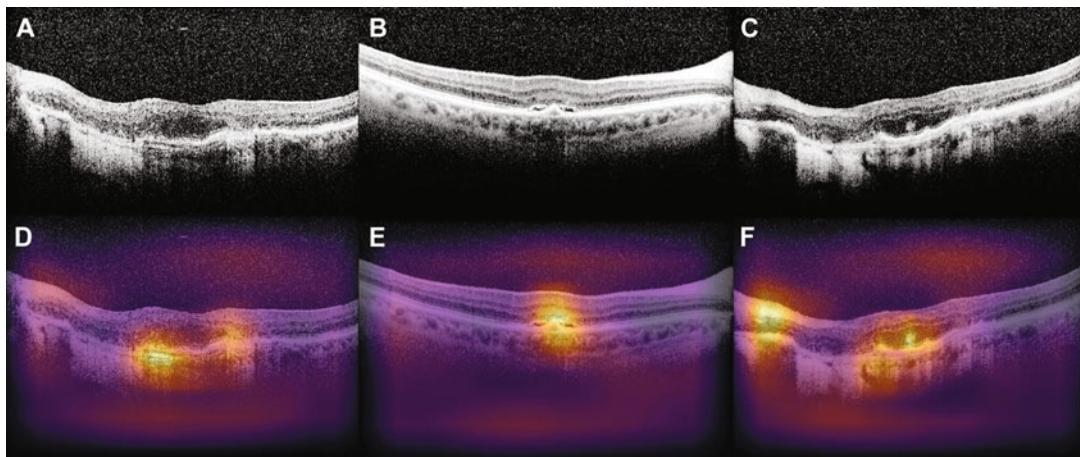


Fig. 2 Examples of identification of age-related macular degeneration (AMD) by deep learning (DL). Optical coherence tomography images in which A, B, and C are used as input images, and heat maps D, E, and F are

identified using the DL algorithm occlusion test. The intensity of the color is determined by the drop in the likelihood of the image being categorized as AMD when this area is occluded [8]

and an area under the ROC curve (AUC) >0.90 , and sensitivities $>90\%$ and specificities $>70\%$ in the ten remaining datasets [12].

The AI models for DR screening in clinical practice may be implemented through a cloud-based configuration, in the office, or built in the retinography cameras. The cloud-based model can integrate AI information into information technology platforms to assist in the analysis of routine images. On the other hand, office-based model implements an application programming interface on computers that integrate the algorithm with retinographies, promoting the image capture followed by an instant diagnosis. In April 2018, the Food and Drug Administration (FDA) approved the first medical device to utilize AI to detect DR, the iDx-DR, which uses a cloud-based algorithm with an almost autonomous retinography camera (Fig. 3). Sufficiently good-quality images can automatically detect vision-threatening and referable diabetic retinopathy [13]. Several studies have been published with the use of AI and detection of DR with retinography images. In one of these studies, [12] 494,661 retinographies were used to develop and validate a DL system for detecting DR and concomitant diseases, such as glaucoma and AMD. The sensitivity and specificity for detecting DR were 91% and 92%, 100% and

91% for high-risk DR, 96% and 87% for glaucoma, and 93% and 89% for AMD, respectively. In this same study, external and multiethnic datasets were tested for DR detection and revealed sensitivities that varied between 92% and 100% and specificities that varied between 73% and 92% [14]. Another study reported a DL algorithm developed to classify the presence of any stage of DR in 75,137 color fundus images from public databases. Their model achieved a 0.97 AUC with a 94% and 98% sensitivity and specificity, respectively. Testing against independent databases achieved AUCs of 0.94 and 0.95, respectively [15].

Glaucoma

Glaucoma is a group of optic neuropathies characterized by progressive degeneration of retinal ganglion cells that have their cell bodies in the inner retina and axons in the optic nerve. Degeneration of these axons results in cupping, a characteristic appearance of the optic disc, and visual field loss [16]. Glaucoma is the main cause of irreversible blindness in the world, with more than 2.9 million blind people in the world [4]. The number of individuals with glaucoma is expected to increase to 111.8 million in 2040 [17].

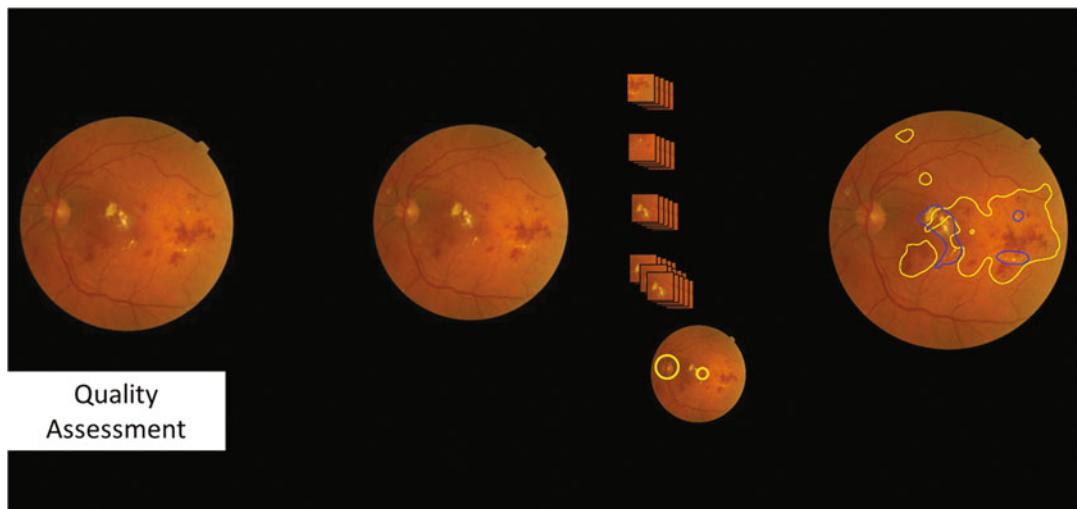


Fig. 3 Working diagram of the iDx-DR algorithm for DR screening. First, a quality assessment indicates whether the image can be used for analysis or whether there are any acquisition artifacts. Then, a DL algorithm using convolutional neural networks evaluates the image for

clinical biomarkers (microaneurysms, hemorrhages, exudates, etc.). As a last step, the evaluation of clinical biomarkers and their location allows for decision-making and classification into no DR, moderate, or high-risk DR [13]

Therefore, early diagnosis of glaucoma is essential, and the use of AI may play a key role in screening, diagnosing, and following the disease.

Major challenges in the development of AI systems for screening or clinical utility in glaucoma include the structural and functional definition of the disease and the heterogeneity of optic nerve findings in glaucoma patients. Training a model capable of diagnosing or detecting early glaucoma is difficult due to three main factors [18]. The first is the presence of borderline parameters between suspicion and normality, such as pre-perimetric glaucoma with little structural change in the peripapillary retinal nerve fiber layer (RNFL) or with a slight thinning of the neuroretinal rim. The second is the lack of standardization of diagnostic criteria for glaucoma, as many studies are based on perimetric criteria. Finally, adequate longitudinal data are needed to identify the suspect cases that will develop glaucoma.

Structural evaluation is essential for the early detection of glaucoma and fundus photographs allow proper documentation and identification of optic disc characteristics at low cost. Ting et al. [12] proposed that a DL algorithm could be

developed to screen for glaucoma with existing teleretinal imaging. Using a large database of 494,661 teleretinal photographs acquired in diabetics, 125,189 of which had been labeled by human graders in the training set, the authors developed an algorithm capable of detecting images that were considered “referable” for glaucoma. In the test dataset, the algorithm detected “referable” glaucoma on photographs with an AUC of 0.942, with a sensitivity of 96.4%, and a specificity of 87.2%. In order to create an approach that would not depend on the low reproducibility of the subjective evaluation of optic disc photographs, Medeiros et al. proposed a DL algorithm that used retinal photographs to estimate RNFL thickness measured with spectral-domain OCT. OCT is a noninvasive imaging test that uses light waves to make cross-section images of the retina, allowing the evaluation and measurement of distinctive layers in a micrometer scale. For glaucoma, the thickness of the RNFL is used as an important metric of damage, widely used to diagnose and detect progression. The dataset included 32,820 pairs of optical disc photographs and OCT scans of the RNFL. There was a strong correlation between the predicted RNFL

value from the DL algorithm's interpretation of the fundus image and the actual RNFL thickness value from the corresponding OCT ($r = 0.832$, $P < 0.001$), with a mean absolute error of approximately 7 μm . The AUC for discriminating healthy from glaucomatous eyes was 0.944. This algorithm would allow the use of a low-cost system to diagnose and stage glaucoma from fundus photographs, which is especially useful where there is no OCT available (Fig. 4a) [19]. Thompson et al. published a follow-up study using a similar approach, in which the OCT Bruch's membrane opening-minimum rim width (BMO-MRW) parameter served as a reference standard for labeling optic disc photographs. BMO-MRW may be particularly useful in images where the optic disc is difficult to grade, such as in eyes with high myopia [20]. The DL predictions were again highly correlated with the actual BMO-MRW values (Pearson's $r = 0.88$, $P < 0.001$), and the AUC for discriminating between glaucomatous and healthy eyes was 0.945 for the DL predictions. Class activation maps confirmed that the neuroretinal rim was

critical to the algorithm's classification (Fig. 4b). A study by Maetschke et al. [21] developed a DL algorithm that could discriminate glaucomatous from healthy eyes using raw, unsegmented OCT volumes of the optic nerve head. The algorithm also performed better than conventional OCT parameters, with AUC of 0.94 compared to 0.89 for a logistic regression model combining OCT parameters. As illustrated in Fig. 4c, the class activation maps appeared to highlight regions in the OCT that have been clinically identified as important to glaucoma diagnosis, particularly the neuroretinal rim, optic disc cupping, and the lamina cribrosa and its surrounding area.

In a subsequent study, Jammal et al. demonstrated that the OCT RNFL predictions from the DL algorithm using fundus photos performed at least as well as and often better than human graders for detecting eyes with reproducible glaucomatous visual field loss. A total of 490 fundus photos of 490 eyes were graded by two glaucoma specialists for the probability of glaucomatous optical neuropathy. Correlations with standard automated perimetry global indices

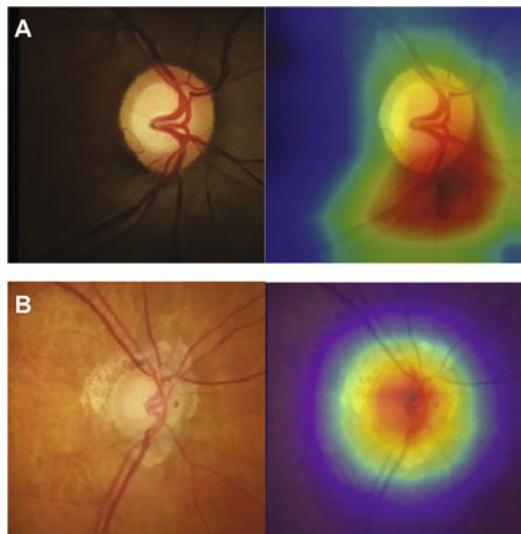
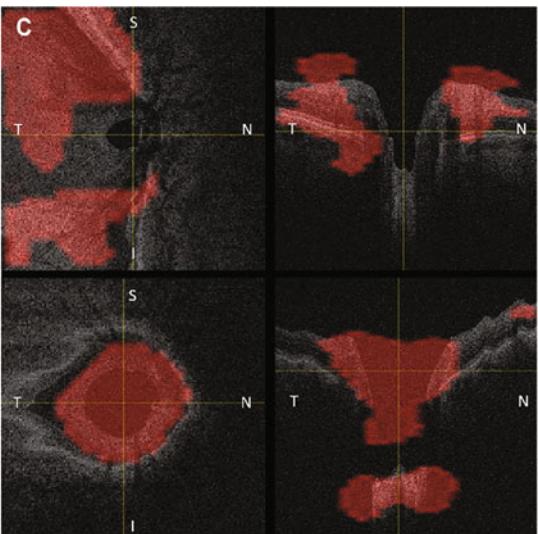


Fig. 4 Class activation maps (CAM) for several examples of deep learning (DL) models. (a) Gradient-weighted CAM from the DL model to predict RNFL thickness from fundus photographs. It can be seen that the heatmap correctly highlights the area of the optic nerve and adjacent RNFL as most relevant for the predictions [19]. (b) Gradient-weighted CAM from the DL model used to



predict rim width in an eye with glaucoma. Note that the heatmap strongly highlights the cup and rim regions [20]. (c) CAM showing the regions in a spectral-domain OCT identified as the most important for the classification of the scan into healthy versus glaucoma eyes. For glaucoma eyes the map highlighted the optic disc cup and neuroretinal rim [21]

were compared between the human gradings and DL RNFL thickness predictions. The DL RNFL thickness predictions had a significantly stronger correlation with mean deviation ($\rho = 0.54$) than the probability of glaucoma given by human graders ($\rho = 0.48$; $P < 0.001$). The partial AUC for the DL algorithm was also significantly higher than that for the probability of glaucoma by human graders (partial AUC = 0.529 vs 0.411, respectively; $P = 0.016$) [22].

Another DL algorithm was developed with training and testing of 1364 glaucoma-compatible fundus photos and 1768 normal fundus photos for the detection of typical glaucomatous signs, such as notching, diffuse thinning of the neural rim, increased disc cupping, RNFL atrophy, disc hemorrhage, and peripapillary atrophy. This algorithm showed an AUC of 0.96 [23]. Another group developed an algorithm using 1758 macular OCT images, incorporating the thickness measurements of the RNFL and ganglion cell complex for the early diagnosis of glaucoma with an AUC of 0.94 [24].

In addition to the analysis of posterior segment OCT, DL models have also been applied to anterior segment OCT images for the diagnosis of angle closure, a sight-threatening event in glaucoma. Fu et al. reported an AUC of 0.96 for a DL system trained to detect angle closure from Visante OCT images, with a sensitivity of 90% and a specificity of 92%, compared to clinician gradings of the same images as the reference standard [25]. In another work, Xu et al. tested three different multiclass CNNs in Chinese-American eyes, and the best-performing classifier detected gonioscopic angle closure with an AUC of 0.93. Given the difficulties associated with the subjective interpretation of gonioscopy and anterior segment OCT images, such models offer great promise in automating the detection of narrow angles [26].

Functional assessment includes the detection of changes in the visual field, which is also fundamental for the diagnosis and management of glaucoma patients. The glaucoma hemifield test (GHT) represents a supervised algorithm that is commonly used in the definition of glaucoma [27]. However, AI techniques can also identify glaucomatous

patterns or detect progression in the visual field, with a capacity equal to or greater than current algorithms. Elze et al. [28] proposed a technique of “archetypal analysis” to classify patterns of visual field loss in glaucoma. The authors showed that the patterns detected by their technique, such as arcuate, partial arcuate, etc., corresponded well to the classification made by human graders. An ML approach called Kalman filtering, which filters out noise from a series of parameter measurements and allows for a trend forecast over time, has been used in longitudinal studies to assess perimeter progression and has shown ability to detect progression up to 2 years before mean deviation linear regression. In this same study, a DL algorithm showed good ability to discriminate between pre-perimetric glaucomatous visual fields and normal visual fields, with an AUC of 0.93. Interestingly, this detection occurred before the appearance of visual field defects defined by Anderson’s criteria [29] (Fig. 5).

The combination of anatomical and functional data has been shown to be superior in relation to these data in isolation in the diagnosis of glaucoma. The ML classifiers allow you to combine these data and assist in the diagnosis of glaucoma. A study has been published to compare the diagnostic ability between ML classifiers and ophthalmologists exposed only to OCT and standard automated perimetry data [30]. The results demonstrated that the diagnostic accuracy of the best ML classifier (the radial base function network, with an AUC of 0.93) was similar to that of glaucoma specialists (AUC = 0.92) and significantly better than general ophthalmologists (AUC = 0.88, $p < 0.05$). These results suggest that the use of ML classifiers can be useful in clinical practice, especially when there is no glaucoma specialist available. Mariottini et al. [31] proposed a combination of global and localized parameters with the requirement for topographic correspondence between structural and functional damage that could be combined in an objective way to be used as a robust reference standard for the development of AI models for glaucoma diagnosis. The criteria assume that the diagnosis of glaucomatous optic neuropathy should involve corresponding structural and functional damage,

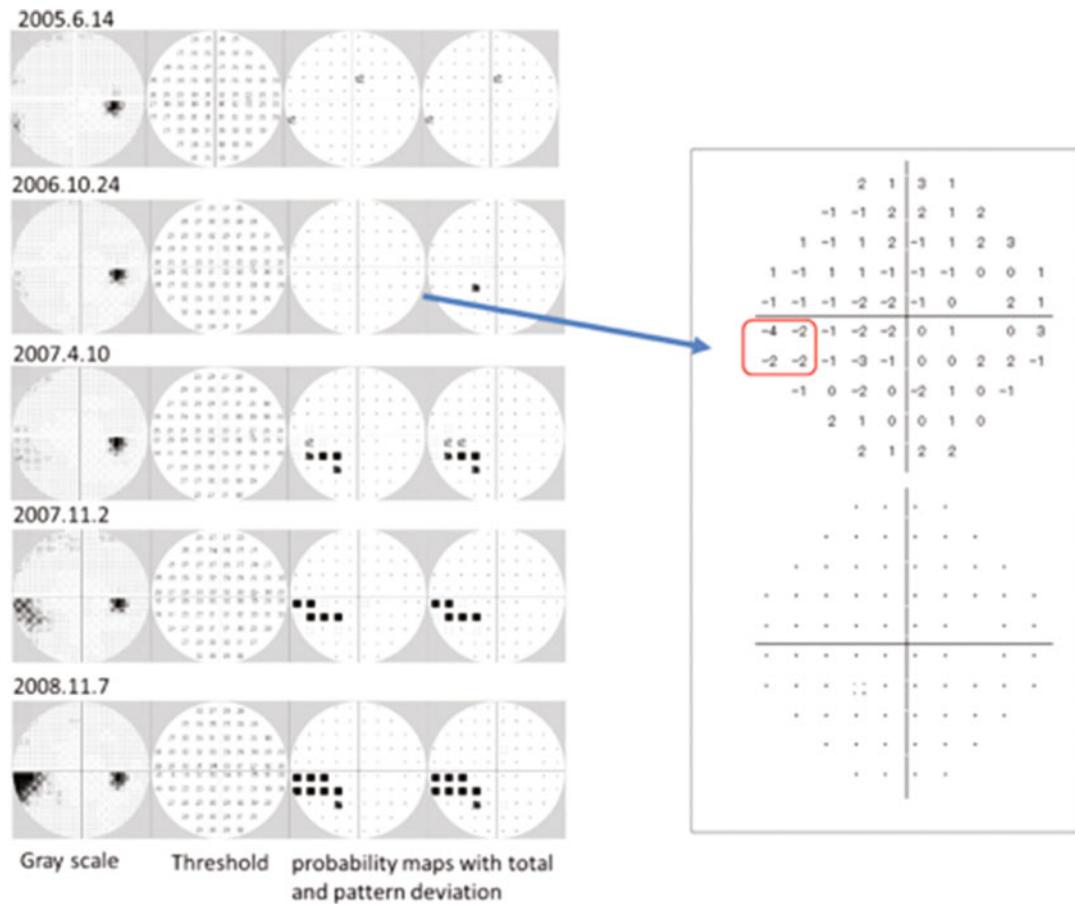


Fig. 5 A case in which the visual field of 10/24/2006 did not show changes that satisfied Anderson's criteria; however, there was a progression that was evident in

subsequent field examinations. With the DL algorithm, the visual field of 10/24/2006 had already been diagnosed as abnormal [29]

based on RNFL assessment by OCT and visual field assessment by standard automated perimetry. The authors then developed a DL model that used fundus photographs to discriminate glaucoma from normal eyes, which had been classified based on the objective reference standard. The model achieved an AUC of 0.92. Of note, an objective reference standard combining OCT and standard automated perimetry data may obviate the need for laborious and time-consuming expert gradings, and may increase the comparability of diagnostic studies across devices and populations.

As for DL applications in detecting progression, there have been only very few studies.

Berchuck et al. [32] proposed a DL variational autoencoder model to learn a low-dimensional representation of standard automated perimetry visual fields using 29,161 fields from 3832 patients. The model was then applied to predict rates of change and future visual field observations. The authors found that at 4 years of follow-up, the model identified 35% of the eyes as progressing versus only 15% for mean deviation. In another study, Park et al. [33] used a recurrent neural network and showed that it achieved better prediction of future visual field observations compared to ordinary least squares linear regression.

Age-Related Macular Degeneration (AMD)

Age-related macular degeneration is a progressive chronic disease of the central retina. Most visual loss occurs in the late stages of the disease due to one of two processes: neovascular (“wet”) age-related macular degeneration and geographic atrophy (“late dry”). In neovascular age-related macular degeneration, choroidal neovascularization breaks through to the neural retina, leaking fluid, lipids, and blood, and leading to fibrous scarring. In geographic atrophy, progressive atrophy of the retinal pigment epithelium, choriocapillaris, and photoreceptors occurs [34].

AMD is a major cause of visual impairment. It is estimated that the number of people with AMD in 2020 will be 196 million and will increase to 288 million in 2040. The treatment of exudative AMD has proved to be effective with the advent of endothelial vascular growth factors inhibitors (anti-VEGF), with a reduction in the incidence of blindness of more than 50% [35]. Patients at higher risk for developing AMD require clinical investigation with imaging tests such as OCT or angiography. AI algorithms can be used as an alternative tool for screening, diagnosis, prognosis, and disease monitoring.

DL-based classification systems have been successfully used to detect and segment AMD lesions (Figs. 6 and 7), and to estimate the risk of progression to advanced stages, or conversion from dry AMD to wet AMD (Fig. 8) [36–38]. A model was created that targets the posterior hyaloid and the epiretinal membrane, which allows an improved assessment of disorders of the vitreomacular interface and retinal pigment epithelium and quantifies the extent of AMD and its atrophic areas. In segmentation, the algorithm outlines the margins of the abnormal areas in OCT scans, allowing the measurement of the area or volume of the abnormal region [38]. Some groups have developed models that allow segmentation of the detachment of the pigment epithelium, delimiting the space between the retinal pigment epithelium and Bruch’s membrane [39].

Since anti-VEGF treatment for AMD involves repeated, invasive, and high-cost application of intravitreal injections over an extended period of time, some authors have applied AI algorithms to predict treatment outcomes and reduce treatment burden. Prahls et al. used a total of 183,402 OCT B-scans from patients with wet AMD to train a DL algorithm to predict whether an injection of anti-VEGF would have to be administered within the next 21 days [40]. Another study used data from the HARBOR clinical trial to develop models to predict the visual acuity of patients who received intravitreal injection of ranibizumab for neovascular AMD and to predict the risk of converting dry AMD to the exudative form [36]. Other AI systems have used the dataset from the age-related eye disease study (AREDS) [41, 42] or their own databases to assess the prognosis of eyes with AMD [37, 38, 40].

Future research is important to assess the generalizability and cost-effectiveness of these AI systems in international multiethnic cohorts. In addition to screening, it will be of great value to generate new algorithms to predict the functional, structural, and treatment results for patients with AMD, with the appropriate stratification of risk profiles.

Conclusions

Given the aging of the population and the increasing financial burden of blindness on health care, there is an urge to innovate in order to improve the detection of eye diseases. AI algorithms are likely to result in a significant shift in ophthalmology in the coming decades, although several challenges need to be resolved to encourage the adoption of AI in health care. First, validation of new diagnostic tests should be based on rigorous methodology. Therefore, the requirements for diagnostic performance may vary considerably depending on whether the algorithm is being considered for community-based or opportunistic screening versus detection or monitoring of disease in a tertiary care center. Secondly, AI approaches require a

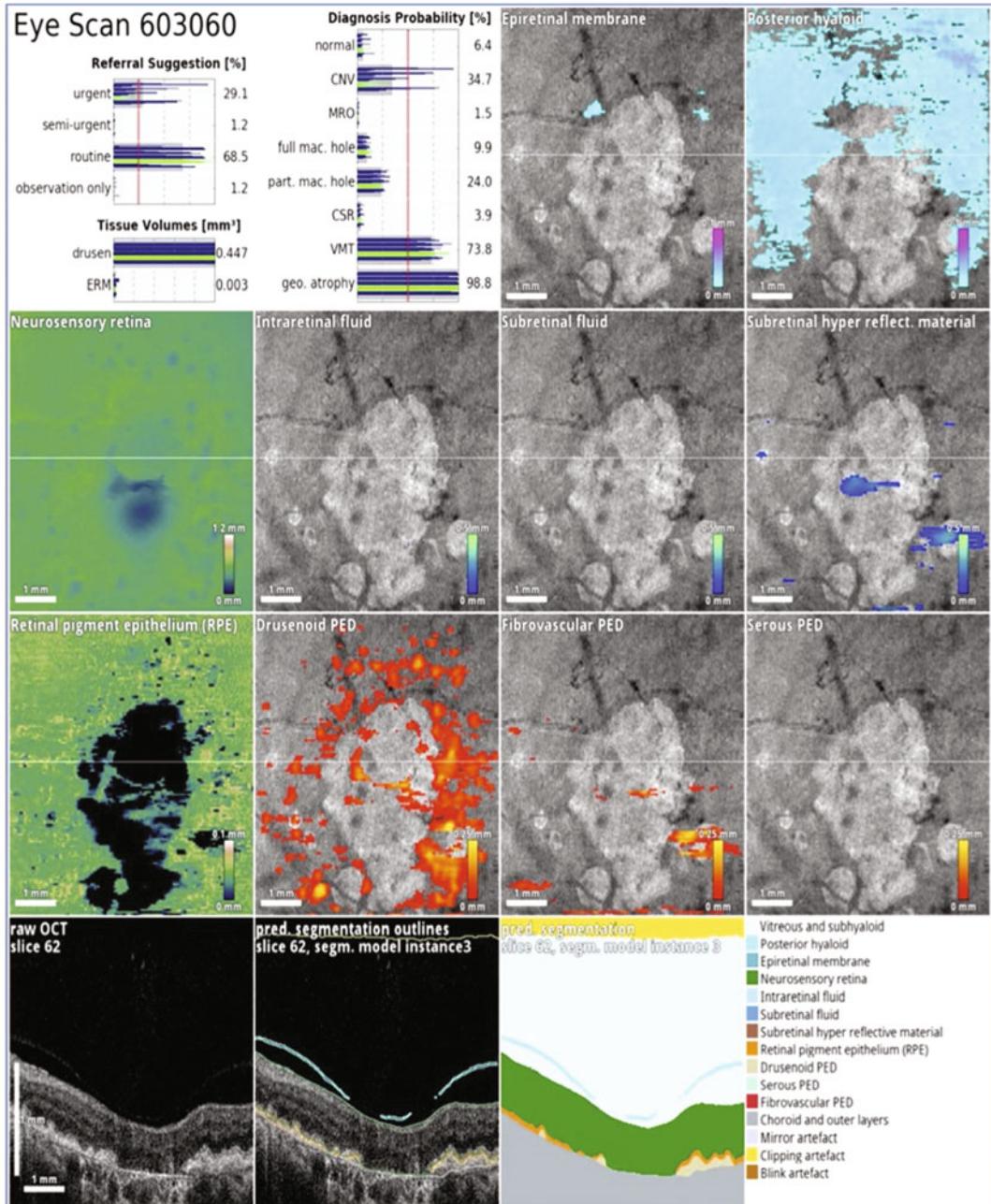


Fig. 6 DL application in the segmentation of OCT images – prototype of the Moorfields DL system (DeepMind). In this case, the system correctly segments the loss of the retinal pigment epithelium (RPE),

highlighting an area of geographic atrophy (red dots in the third line). Geographic atrophy is surrounded by foci of pigmented drusenoid detachment. The partial detachment of the posterior hyaloid is also outlined in the last row [37]

large number of images. However, increasing data does not necessarily improve the performance of a network, which will depend on the

quality of the images, the reference standard, and the representativeness of the data. A possible solution includes the sharing of data between

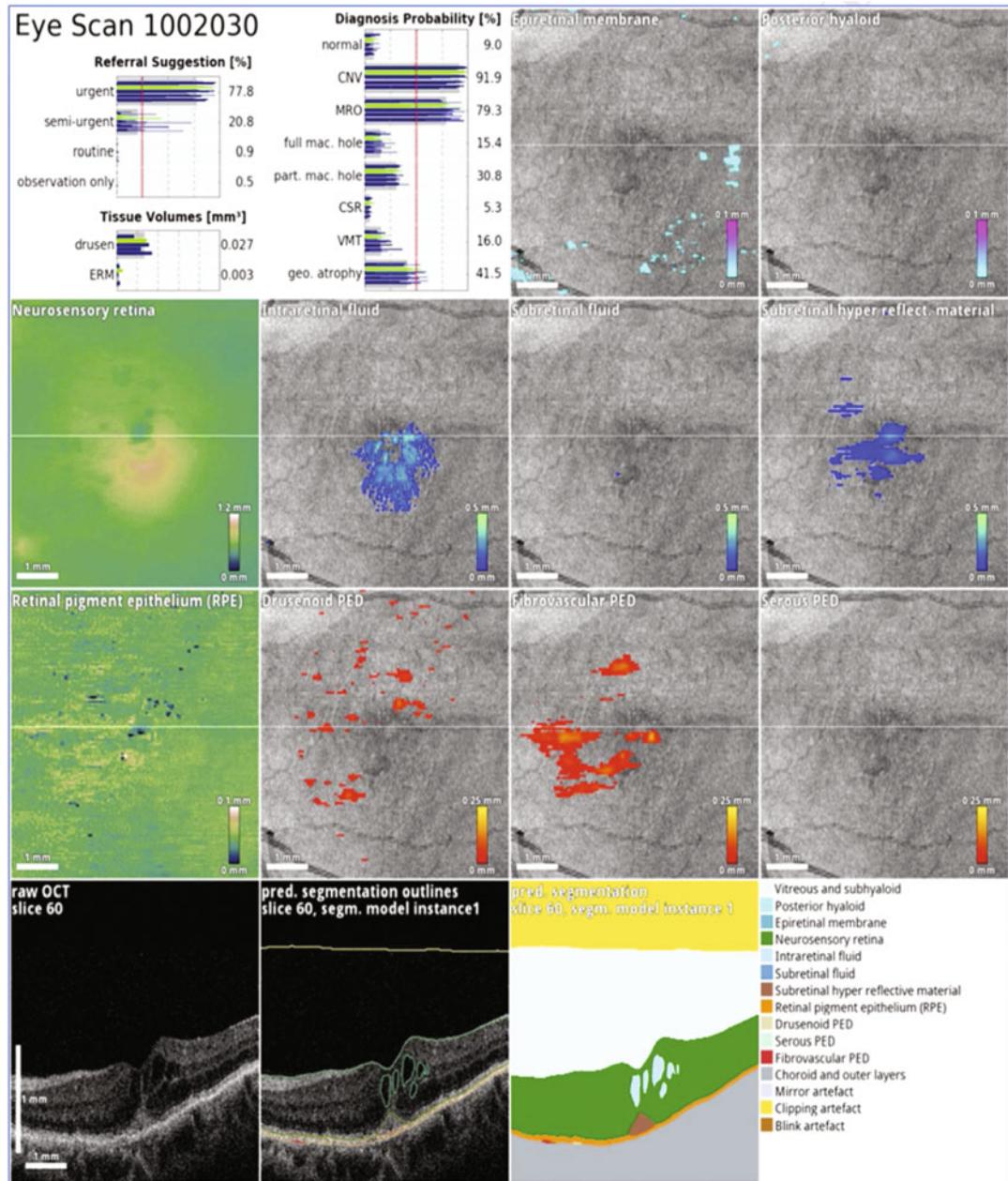


Fig. 7 DL application in the segmentation of OCT images with DeepMind. In the case of angiomatic proliferation of the retina, the system correctly segments an area of intraretinal fluid superimposed on an area of subretinal hyper-reflective material. The presence of macular edema

and choroidal neovascularization was detected and a referral to the specialist was suggested. Through the creation of a tissue representation on a 2D map (seen below and on the right), the system facilitates interpretation by the ophthalmologist [37]

different research centers, always privileging ethical and privacy standards. Data sharing also requires technical solutions, including data

storage, management, and analysis, which implies high investments in hardware and software, and specialized labor.

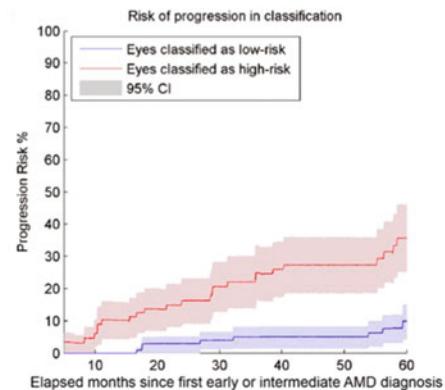
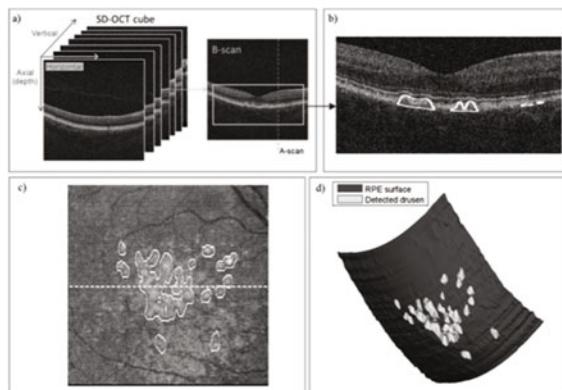


Fig. 8 Artificial intelligence to predict risk of progression of AMD from a set of quantitative parameters extracted from drusen. The system assigns each patient a risk rate and

In conclusion, AI models have consistently shown robust performance in detecting and staging a number of eye diseases using standard imaging tests in ophthalmology. Although many of these algorithms are not yet available for commercial use, it is expected that AI will soon be able to aid ophthalmologists to provide the best and effective clinical care. If eye care professionals intend to maintain control of their occupational future, they will need to understand and incorporate intelligent algorithms into practice.

is highly accurate, as shown in the Kaplan-Meier curve on the right [36]

References

- Klein R, Klein BE. The prevalence of age-related eye diseases and visual impairment in aging: current estimates. *Invest Ophthalmol Vis Sci*. 2013;54(14): ORSF5-13. <https://doi.org/10.1167/iovs.13-12789>.
- Ting DSW, Cheung GCM, Wong TY. Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review. *Clin Exp Ophthalmol*. 2016;44(4):260–77. <https://doi.org/10.1111/ceo.12696>.
- Chua J, Baskaran M, Ong PG, Zheng Y, Wong TY, Aung T, et al. Prevalence, risk factors, and visual features of undiagnosed glaucoma: the Singapore epidemiology of eye diseases study. *JAMA Ophthalmol*. 2015;133(8):938–46. <https://doi.org/10.1001/jamaophthalmol.2015.1478>.
- Flaxman SR, Bourne RRA, Resnikoff S, Ackland P, Braithwaite T, Cicinelli MV, et al. Global causes of blindness and distance vision impairment 1990–2020: a systematic review and meta-analysis. *Lancet Glob Health*. 2017;5(12):e1221–e34. [https://doi.org/10.1016/s2214-109x\(17\)30393-5](https://doi.org/10.1016/s2214-109x(17)30393-5).
- Wheatley CM, Dickinson JL, Mackey DA, Craig JE, Sale MM. Retinopathy of prematurity: recent advances in our understanding. *Arch Dis Child Fetal Neonatal Ed*. 2002;87(2):F78–82. <https://doi.org/10.1136/fn.87.2.f78>.
- Molina-Casado JM, Carmona EJ, García-Feijoó J. Fast detection of the main anatomical structures in digital retinal images based on intra- and inter-structure relational knowledge. *Comput Methods Prog Biomed*. 2017;149:55–68. <https://doi.org/10.1016/j.cmpb.2017.06.022>.
- Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*. 2018;2(3):158–64. <https://doi.org/10.1038/s41551-018-0195-0>.
- Lee CS, Baughman DM, Lee AY. Deep learning is effective for classifying Normal versus age-related macular degeneration OCT images. *Ophthalmol Retina*. 2017;1(4):322–7. <https://doi.org/10.1016/j.oret.2016.12.009>.
- Cheung N, Mitchell P, Wong TY. Diabetic retinopathy. *Lancet*. 2010;376(9735):124–36. [https://doi.org/10.1016/S0140-6736\(09\)62124-3](https://doi.org/10.1016/S0140-6736(09)62124-3).
- Jeganathan VSE, Wang JJ, Wong TY. Ocular associations of diabetes other than diabetic retinopathy. *Diabetes Care*. 2008;31(9):1905–12.
- Krause J, Gulshan V, Rahimy E, Karth P, Widner K, Corrado GS, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*. 2018;125(8):1264–72. <https://doi.org/10.1016/j.ophtha.2018.01.034>.
- Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*. 2017;318(22): 2211–23. <https://doi.org/10.1001/jama.2017.18152>.

13. Abramoff MD, Lou Y, Erginay A, Clarida W, Amelon R, Folk JC, et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Invest Ophthalmol Vis Sci.* 2016;57(13):5200–6. <https://doi.org/10.1167/iovs.16-19964>.
14. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA.* 2016;316(22):2402–10. <https://doi.org/10.1001/jama.2016.17216>.
15. Gargya R, Leng T. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology.* 2017;124(7):962–9. <https://doi.org/10.1016/j.ophtha.2017.02.008>.
16. Weinreb RN, Khaw PT. Primary open-angle glaucoma. *Lancet.* 2004;363(9422):1711–20. [https://doi.org/10.1016/s0140-6736\(04\)16257-0](https://doi.org/10.1016/s0140-6736(04)16257-0).
17. Tham YC, Li X, Wong TY, Quigley HA, Aung T, Cheng CY. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology.* 2014;121(11):2081–90. <https://doi.org/10.1016/j.ophtha.2014.05.013>.
18. Chang RT, Singh K. Glaucoma suspect: diagnosis and management. *Asia Pac J Ophthalmol (Phila).* 2016;5(1):32–7. <https://doi.org/10.1097/ajop.0000000000000173>.
19. Medeiros FA, Jammal AA, Thompson AC. From machine to machine: an OCT-trained deep learning algorithm for objective quantification of glaucomatous damage in fundus photographs. *Ophthalmology.* 2019;126(4):513–21. <https://doi.org/10.1016/j.ophtha.2018.12.033>.
20. Thompson AC, Jammal AA, Berchuck SI, Mariottini EB, Medeiros FA. Assessment of a segmentation-free deep learning algorithm for diagnosing glaucoma from optical coherence tomography scans. *JAMA Ophthalmol.* 2020. <https://doi.org/10.1001/jamaophthalmol.2019.5983>.
21. Maetschke S, Antony B, Ishikawa H, Wollstein G, Schuman J, Garnavi R. A feature agnostic approach for glaucoma detection in OCT volumes. *PLoS One.* 2019;14(7):e0219126. <https://doi.org/10.1371/journal.pone.0219126>.
22. Jammal AA, Thompson AC, Mariottini EB, Berchuck SI, Urata CN, Estrela T, et al. Human versus machine: comparing a deep learning algorithm to human gradings for detecting glaucoma on fundus photographs. *Am J Ophthalmol.* 2019. <https://doi.org/10.1016/j.ajo.2019.11.006>.
23. Shibata N, Tanito M, Mitsuhashi K, Fujino Y, Matsura M, Murata H, et al. Development of a deep residual learning algorithm to screen for glaucoma from fundus photography. *Sci Rep.* 2018;8(1):14665. <https://doi.org/10.1038/s41598-018-33013-w>.
24. Asaoka R, Murata H, Hirasawa K, Fujino Y, Matsura M, Miki A, et al. Using deep learning and transfer learning to accurately diagnose early-onset glaucoma from macular optical coherence tomography images. *Am J Ophthalmol.* 2019;198:136–45. <https://doi.org/10.1016/j.ajo.2018.10.007>.
25. Fu H, Baskaran M, Xu Y, Lin S, Wong DWK, Liu J, et al. A deep learning system for automated angle-closure detection in anterior segment optical coherence tomography images. *Am J Ophthalmol.* 2019;203:37–45. <https://doi.org/10.1016/j.ajo.2019.02.028>.
26. Xu BY, Chiang M, Chaudhary S, Kulkarni S, Pardeshi AA, Varma R. Deep learning classifiers for automated detection of gonioscopic angle closure based on anterior segment OCT images. *Am J Ophthalmol.* 2019;208:273–80. <https://doi.org/10.1016/j.ajo.2019.08.004>.
27. Åsman P, Heijl A. Glaucoma Hemifield test: automated visual field evaluation. *Arch Ophthalmol.* 1992;110(6):812–9. <https://doi.org/10.1001/archophht.1992.01080180084033>.
28. Elze T, Pasquale LR, Shen LQ, Chen TC, Wiggs JL, Bex PJ. Patterns of functional vision loss in glaucoma determined with archetypal analysis. *J R Soc Interface.* 2015;12(103). <https://doi.org/10.1098/rsif.2014.1118>.
29. Asaoka R, Murata H, Iwase A, Araie M. Detecting preperimetric glaucoma with standard automated perimetry using a deep learning classifier. *Ophthalmology.* 2016;123(9):1974–80. <https://doi.org/10.1016/j.ophtha.2016.05.029>.
30. Shigueoka LS, Vasconcellos JPC, Schiimiti RB, Reis ASC, Oliveira GO, Gomi ES, et al. Automated algorithms combining structure and function outperform general ophthalmologists in diagnosing glaucoma. *PLoS One.* 2018;13(12):e0207784. <https://doi.org/10.1371/journal.pone.0207784>.
31. Mariottini EB, Datta S, Dov D, Jammal AA, Berchuck SI, Tavares IM, et al. Artificial intelligence mapping of structure to function in glaucoma. *Transl Vis Sci Technol.* 2020;9(2):19. <https://doi.org/10.1167/tvst.9.2.19>.
32. Berchuck SI, Mukherjee S, Medeiros FA. Estimating rates of progression and predicting future visual fields in glaucoma using a deep variational autoencoder. *Sci Rep.* 2019;9(1):18113. <https://doi.org/10.1038/s41598-019-54653-6>.
33. Park K, Kim J, Lee J. Visual field prediction using recurrent neural network. *Sci Rep.* 2019;9(1):8385. <https://doi.org/10.1038/s41598-019-44852-6>.
34. Lim LS, Mitchell P, Seddon JM, Holz FG, Wong TY. Age-related macular degeneration. *Lancet.* 2012;379(9827):1728–38. [https://doi.org/10.1016/S0140-6736\(12\)60282-7](https://doi.org/10.1016/S0140-6736(12)60282-7).
35. Bressler NM, Doan QV, Varma R, Lee PP, Suñer IJ, Dolan C, et al. Estimated cases of legal blindness and visual impairment avoided using ranibizumab for choroidal neovascularization: non-Hispanic white population in the United States with age-related macular degeneration. *Arch Ophthalmol.* 2011;129(6):709–17. <https://doi.org/10.1001/archophthalmol.2011.140>.
36. Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, Waldstein SM, Bogunović H. Artificial intelligence in

- retina. *Prog Retin Eye Res.* 2018;67:1–29. <https://doi.org/10.1016/j.preteyeres.2018.07.004>.
37. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med.* 2018;24(9):1342–50. <https://doi.org/10.1038/s41591-018-0107-6>.
38. Lee CS, Tyring AJ, Deruyter NP, Wu Y, Rokem A, Lee AY. Deep-learning based, automated segmentation of macular edema in optical coherence tomography. *Biomed Opt Express.* 2017;8(7):3440–8. <https://doi.org/10.1364/boe.8.003440>.
39. Xu Y, Yan K, Kim J, Wang X, Li C, Su L, et al. Dual-stage deep learning framework for pigment epithelium detachment segmentation in polypoidal choroidal vasculopathy. *Biomed Opt Express.* 2017;8(9):4061–76. <https://doi.org/10.1364/BOE.8.004061>.
40. Prahs P, Radeck V, Mayer C, Cvetkov Y, Cvetkova N, Helbig H, et al. OCT-based deep learning algorithm for the evaluation of treatment indication with anti-vascular endothelial growth factor medications. *Graefes Arch Clin Exp Ophthalmol.* 2018;256(1):91–8. <https://doi.org/10.1007/s00417-017-3839-y>.
41. Burlina P, Pacheco KD, Joshi N, Freund DE, Bressler NM. Comparing humans and deep learning performance for grading AMD: a study in using universal deep features and transfer learning for automated AMD analysis. *Comput Biol Med.* 2017;82:80–6.
42. Grassmann F, Mengelkamp J, Brandl C, Harsch S, Zimmermann ME, Linkohr B, et al. A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. *Ophthalmology.* 2018;125(9):1410–20. <https://doi.org/10.1016/j.ophtha.2018.02.037>.



Kevin Hilbert

Contents

Introduction	1568
AI and Machine Learning for Precision Medicine in Depression and Anxiety Disorders	1568
Aims for Using AI in Depression and Anxiety	1570
Risk	1570
Diagnosis	1571
Treatment Outcome	1572
Suicidality	1574
Relapse	1574
Outlook	1575
Cross-References	1576
References	1576

Abstract

Depression and anxiety disorders are common and associated with considerable impairment. Although effective treatments are available, many patients do not respond sufficiently. By matching treatments to individual patient characteristics, precision medicine bears great potential for improving detection and treatment of these disorders, and AI and machine learning may be the most important tools for realizing this potential. By learning from data

with already known outcomes, predictive models can be trained which, after careful evaluation, may be used to predict unknown outcomes for individual patients in clinical practice, such as disorder onset or treatment success. For depression and anxiety disorders, a variety of potential applications have been investigated, including risk and onset prediction, (differential) diagnosis, treatment outcome prediction and treatment selection, prediction of suicidality, and relapse prediction. Many of these studies reported very promising findings, but so far, few replications or predictions in independent samples are available. Additionally, prediction performance has to be further optimized before clinical application. Rigorous research is needed in

K. Hilbert (✉)
Department of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany
e-mail: kevin.hilbert@hu-berlin.de

order to replicate, validate, and characterize promising predictive models in the future.

Keywords

Classification · Prediction · Clustering · Machine learning · Artificial intelligence · Biomarker · Risk · Diagnosis · Outcome · Relapse

Introduction

Depression and anxiety disorders are among the most common mental disorders [1]. Recent epidemiologic estimates suggest that 30% of all individuals will have a diagnosis of depression (i.e., a unipolar depressive disorder such as major depression or dysthymia) at some point in their life and about 40% will have one of an anxiety disorder [2]. Patients with these disorders are also frequent treatment utilizers, placing both diagnostic groups among the most common disorders seen in outpatient and (for depression) inpatient settings for mental disorders. Both clinical depression and anxiety are associated with considerable individual [3] and societal burden [4]. This includes a considerable reduction in life expectancy, in the range of 7–11 years and thus comparable to moderate-to-heavy smoking, and an about 20-fold risk for suicide, both for depression [5]. For anxiety disorders, the increase in mortality is less pronounced, but still these disorders are associated with an odds ratio >3 for suicide compared to the general population [5].

For depression and anxiety disorders, a number of effective treatment options are available. The main options for depression are pharmacological treatment, including selective serotonin reuptake inhibitors (SSRIs), serotonin–norepinephrine reuptake inhibitors (SNRIs), monoamine oxidase inhibitors (MAOIs), and tricyclic antidepressants as the most often used medications, and psychotherapy, including cognitive behavioral therapy (CBT), interpersonal therapy (IPT), and psychodynamic therapy as most common therapies. Further treatments are available such as electroconvulsive therapy, transcranial magnetic stimulation, and others, and are particularly applied in the case of

treatment-resistant depression. For anxiety disorders, psychotherapies are regarded as first-line treatment. The best evidence is available for CBT, but other therapies such as mindfulness-based therapies and psychodynamic therapy are used as well. Pharmacological treatment is also common, again including SSRIs, SNRIs, and others, but also benzodiazepines as acute treatment.

Although effective treatments for depression and anxiety are available, there is still a substantial portion of patients who do not respond sufficiently [6–8]. As such, patients with these disorders may have much to gain from precision medicine, which aims to match treatments to individual patient characteristics [9]. Artificial intelligence is among the most promising tools in order to realize this potential. This chapter aims to summarize the main research questions that have so far been investigated using artificial intelligence in the field, describe selected studies and outcomes, and present an outlook for the future.

AI and Machine Learning for Precision Medicine in Depression and Anxiety Disorders

The application of AI for depression and anxiety disorders is realized by the use of a specific type of machine learning, supervised learning. Basically, these methods extract the relationship between a set of given variables, called features, and a set of outcomes (labels) from known data. Importantly, this relationship is not explicitly given or defined by the researcher, but identified from the data by the algorithm, hence “machine learning.” After the relationship has been established, the algorithm can be applied to make predictions for new data, for which the outcomes are not yet known. However, before doing so, it is crucial that the algorithm is applied on independent data in order to evaluate its performance. This is necessary as the established association may not generalize well beyond the initial dataset for a number of potential reasons, such as the algorithm overly depending on noise in the initial dataset, or the initial dataset being not representative of the underlying population. The process of

establishing the association between input data and outcomes is commonly referred to as training and the process of evaluating the algorithm performance as testing. Figure 1 provides a schematic overview on the generation, evaluation, and use of predictive models via machine learning. The outlined concepts may be easily illustrated by one example from the field: imagine a patient with treatment-resistant depression. Before admission to another inpatient treatment program, the patient and clinicians would like to know whether this new program is promising. Here, a machine learning algorithm could be trained to predict response or nonresponse based on the patient's characteristics before initializing treatment. Pretreatment characteristics and treatment outcomes from previous patients in the clinic would constitute the training data. If an algorithm can be developed and shown to predict treatment outcome with sufficient accuracy, it could then be used on the patient in question, for whom the treatment outcome is not yet known. However, testing the algorithm before application in real clinical decision-making is critical in order to evaluate its generalizability to new data. This could for instance be limited if the treatment program changed between earlier patients comprising the training data and the patient in question, or if by chance only women have been included in the training data, but the patient is male.

This simple example demonstrates that the use of AI and machine learning for depression and anxiety is a complex endeavor, but it also points towards its great potential: if successful, it may be

a valuable tool supporting clinical decision-making by providing evidence-based predictions for individual patients. Additionally, machine learning models particularly excel if the outcome depends on complex interactions between a variety of variables, which is commonly the case for all kinds of clinical outcomes in the field of mental health. Finally, the machine learning algorithm builds the predictive model from the data itself with no need for the researcher or clinician to specify a certain theoretical model, which is beneficial if the theoretical connections are not yet fully understood. Outcomes can either be categorial (e.g., therapy response or nonresponse, presence or absence of a specific diagnosis), in which case the performance of the model is often evaluated in the field by examining sensitivity, specificity, overall accuracy, or area under the curve (AUC) of a receiver operating characteristic curve, or dimensional (e.g., symptom severity), in which case the model performance is often measured by quantifying the difference between true and predicted scores. Typical indices are the mean absolute error (MAE) or the root mean square error (RMSE). For optimizing the performance, a plethora of machine learning algorithms with differing advantages can be used, the hyperparameters controlling how these algorithms work can be tuned, features can be engineered, tuned, or selected to find the best predictors, and the sample size available for training and testing can be increased – models based upon larger samples usually work better until a plateau of maximum performance has been reached.

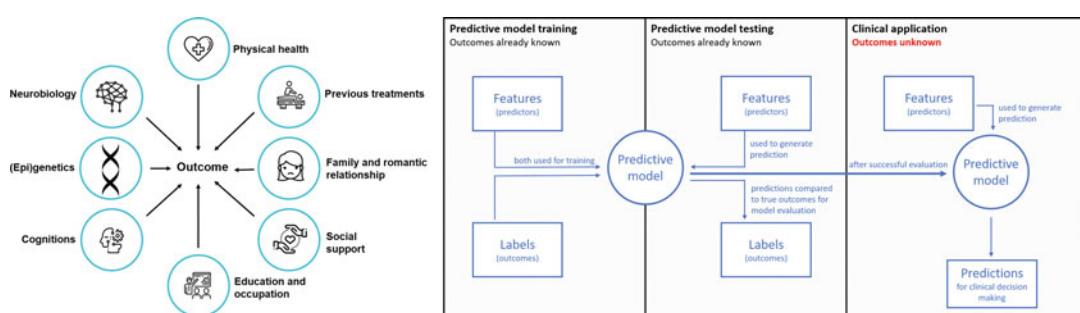


Fig. 1 A non-exhaustive selection of potential data sources for outcome prediction in depression and anxiety disorders (left side) and a schematic overview on the

process of training, testing and application of predictive models via machine learning (right side)

While the remainder of this chapter exclusively reports on findings made by supervised learning, i.e., where the outcome categories have been given by the researcher, it should be noted that there is a second type of machine learning called unsupervised learning which has also been applied to depression and anxiety disorders. In unsupervised learning, labels are not given, but the machine learning algorithm aims to find patterns in the data on its own. Common examples for unsupervised machine learning are clustering methods. In recent years, some studies employed sophisticated unsupervised learning on clinical [10] and neuroimaging data [11] in depression. The approach is interesting as it may be possible to detect subtypes associated with different disorder-courses or coming from separate etiological pathways, but as this research is still very scarce, it will not be presented in more detail here.

Aims for Using AI in Depression and Anxiety

When utilizing machine learning for depression and anxiety disorders, precision medicine can be understood more broadly, including not only treatment selection in the strict sense but various kinds of clinical decision-making based on

evidence-based predictions for individual patients. This may even include decision-making for individuals who are not yet affected by a disorder, for instance when deciding on access for indicated prevention programs based on individual risk scores. On the other end of the spectrum, it may also include decisions after treatment termination, for instance based on relapse predictions. Figure 2 uses a simplified sketch of depression symptoms before and over the course of an episode in order to illustrate stages of the disorder and associated clinical decisions which may be supported by AI in the future. Of course, the same stages and clinical decisions apply to anxiety disorders. In the following sections, exemplary studies for each field of application are presented, particular challenges or approaches are noted, and the case for AI and machine learning is illustrated. Notably, only a selection of studies can be discussed, as this is a fast-evolving field of study.

Risk

Substantial research suggests that individuals affected by mental disorders experience substantial delay before treatment is initiated, which is potentially related to individual and care-system barriers and considerable stigma still associated

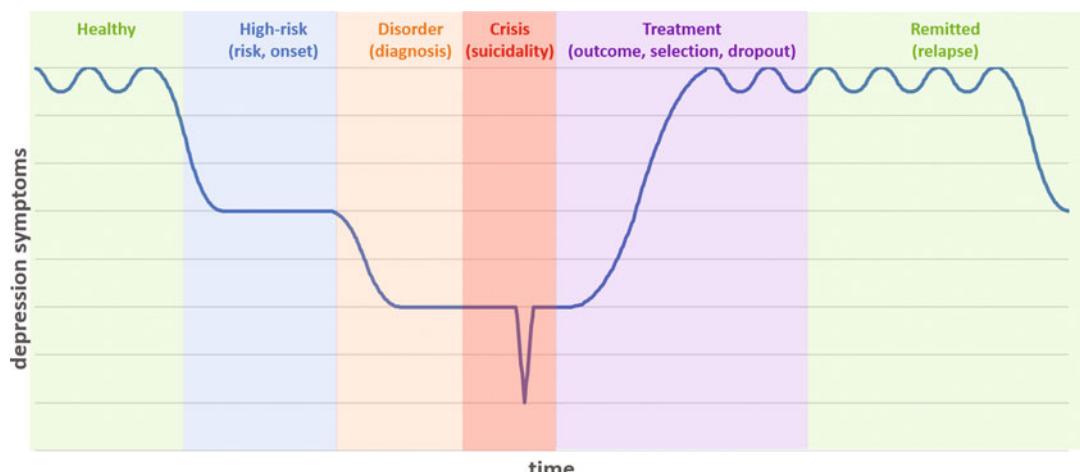


Fig. 2 The figure depicts change in depression symptoms over time, starting with day-to-day mood changes before the onset of the disorder. AI may support clinical decision

making at many different stages, illustrated by coloured boxes. Within parentheses, exemplary targets for classification and prediction via machine learning are given

with mental disorders [12]. Treatment delay is also common for depression and anxiety disorders. Against this background, individual risk prediction may be highly advantageous for identifying and monitoring high-risk individuals in order to shorten time to intervention. Additionally, prevention programs might be easier distributed to the relevant target groups. For depression and anxiety disorders specifically, a number of interesting studies have been published, but particularly for anxiety disorders there is considerably fewer data available compared to the prediction of psychosis onset or PTSD onset after trauma.

Of particular interest are two papers that built predictive models based on data from large nationwide prospective surveys, in the USA and the Republic of Korea, respectively [13, 14]. The first paper examined prediction for a range of disorders, including depression, social anxiety disorder, panic disorder, and generalized anxiety disorder, while the second study only investigated the onset of depression. Additionally, one notable difference between the studies is that the depression diagnosis was based on structured clinical interviews in one study [13], while based on cut-off scores in the other [14]. Both studies reported very encouraging prediction performances: AUCs in the first study ranged between 0.75 and 0.80, depending on the specific anxiety or depressive disorder [13], while an AUC = 0.87 was given in the second study [14]. Interestingly, the most relevant predictors differed, with clinical and biographic data such as previous PTSD comorbidity, trauma, and childhood adversities scoring high for study one [13] and satisfaction within a number of areas such as health and significant social relationships being the most important predictors in study two [14].

Given that depression and anxiety disorders often begin before adulthood, onset prediction in corresponding samples is particularly interesting. A paper by Cohen and colleagues used data from two prospective studies in youth to predict depression onset [15]. Here, a particular merit was that the predictive model was trained on one of the samples but tested on the other, i.e., the predictors had to generalize on a completely independent

sample. When using a cumulative risk score based on rumination, impairment, and negative affect in a regression approach the authors found AUCs for depression onset of 0.73 for the training sample and 0.68 for the independent test sample [15]. First studies in further high-risk cohorts such as elderly individuals or postpartum women are also emerging [16, 17].

Diagnosis

Under this umbrella section a number of studies can be sorted that investigated the use of AI and machine learning for diagnostic purposes in a broader sense, i.e., for classifying individuals with depression or anxiety disorders versus healthy comparison subjects or for classifying individuals with depression or anxiety versus other disorders. Generally, the classification of individuals with these disorders versus healthy individuals seems comparably simple, as many studies reported very high classification accuracies for this task, often exceeding 80–90% [18]. However, it should be noted that these studies often operated on small sample sizes, which may lead to overly optimistic estimates of classification performance [19]. Consequently, some reviews reported that classification accuracy decreases with increasing sample size [18, 19]. In a similar vein, a large-scale study that aimed to predict trait anxiety composite scores in a second sample unrelated to the sample where the model was trained found that the predictive model based on structural and functional neuroimaging data failed to exceed chance level [20]. Notably, trait anxiety is different although related to an anxiety disorder diagnosis, but still this methodologically well-conducted study suggests that generalization to truly independent test samples may be difficult despite promising model performance during cross-validation in the original sample.

As distinguishing individuals affected by depression and anxiety disorders from healthy individuals can usually be achieved reliably with the clinical instruments already available, supporting differential diagnosis (i.e., classifying

between several potential diagnoses) may be the more interesting case for AI and machine learning. Here, two example cases are considered: first, a current depressive episode presented by a patient is likely part of a unipolar depressive disorder, but it may also be part of bipolar disorder. Rapidly identifying those patients with bipolar disorder would be highly helpful for treatment selection, thus an AI solution might yield high clinical benefit. Second, some disorders within the internalizing spectrum such as major depression and generalized anxiety disorder (GAD) share not only a variety of symptoms and diagnostic criteria, but also risk factors or descriptive features. Again, this may hamper diagnosis and treatment selection and thus make an interesting use case for AI and machine learning. For the classification of unipolar depressive versus bipolar patients, early studies in comparably small samples found high classification accuracies of 90% and 81% for functional [21] and structural [22] MRI data. Also using neuroimaging data acquired with near-infrared spectroscopy, Takizawa and colleagues classified more than 400 matched patients with current depressive symptoms but an underlying diagnosis of major depression, bipolar disorder, or schizophrenia that were collected at seven different sites [23]. Optimal thresholds for binary classifications were determined on data from one site and then tested on data from the remaining six sites, achieving AUCs of 0.74 for depression versus bipolar disorder and 0.86 for depression versus schizophrenia. While these accuracies are somewhat below those from initial studies, they were achieved in considerably larger and more heterogeneous multisite data. These promising results are complemented by a very recent work using routinely acquired outpatient data in order to predict transition to bipolar disorder within 3 months after initializing depression treatment with antidepressants [24]. Several models were trained in the largest part of a first cohort comprising more than 40,000 patients in total, and then tested on a heldout set from this first cohort as well as on a completely independent second cohort of more than 25,000 patients. The best performing models achieved AUCs of 0.76 in the heldout data from

the first cohort and of 0.70 in the independent second cohort [24]. With its very large sample and independent out-of-site testing, its reliance on routinely acquired outpatient data and its naturalistic setting with very imbalanced groups (i.e., very few patients transitioning to bipolar disorder), this final study shows outstanding internal and external validity. For the classification of unipolar depressive versus other internalizing patients, an example study used clinical questionnaire, structural MRI, and cortisol data to separate individuals with GAD and major depression from healthy individuals and then to separate the individuals with major depression but without GAD from those with both disorders or GAD only [25]. In this study, clinical questionnaire data alone was sufficient and suitable to separate healthy individuals from individuals with a disorder, but only MRI and cortisol data was suitable to separate the diagnostic groups. When combining all multimodal data types, significant classification models for both tasks could be constructed [25]. Despite its limited sample size, this study demonstrates that some data types may be more suitable for some tasks than others, and provides an interesting case for the combination of multimodal data using machine learning.

As the selection of studies presented here demonstrates, a substantial proportion of studies in the literature used neuroimaging data for these classification tasks. However, also other modalities are increasingly investigated, with some of the most intriguing approaches using data that can be collected with no additional effort in real life, such as passive smartphone data, wearables, or social media postings. In the long term, this may enable identifying affected individuals that have not yet established contact to the health system very quickly and thus may lead to shortened delay until treatment is initiated.

Treatment Outcome

The prediction of treatment outcomes based on individual patient information is at the core of precision medicine. For depression and anxiety disorders, an increasing number of studies

becomes available, with many investigations reporting promising and encouraging results. Still, findings for the prediction of treatment outcomes are often based on very distinct predictive models, with few conceptual replications or comparable approaches across studies. In order to provide a concise overview, this section will point out three recurring themes in the literature, starting with the emergence of comparably large-scale studies for binary outcome prediction across a number of settings and samples, then outlining the use of the personalized advantage index (PAI; [26]) for predicting superiority of one treatment over another for an individual patient, and finally describing the repeated finding of anterior cingulate cortex-based neuroimaging predictors that achieve good prediction performances in depression and anxiety disorders.

There are a number of binary outcome studies (e.g., remitter or responder vs. nonremitter or nonresponder) in samples of considerable size available. A first study trained a model to predict remission after SSRI treatment in a sample of approx. 2000 patients and reported prediction performance in several treatment groups of an independent sample [27]. Interestingly, the predictive model's capacity to generalize was dependent on the antidepressant type: accuracies of 59.6% and 59.7% were achieved when predicting outcomes in groups treated with SSRI alone or in combination in the independent sample, while only a non-significant accuracy of 51% was achieved for a group treated with antidepressants different from SSRIs. Comparable prediction performances were reported in a study predicting treatment dropout across two large treatment centers totaling more than 20,000 patients for model training and more than 15,000 patients for testing [28]. Here, prediction resulted in a mean AUC of 0.68 in the independent test sample for the best performing approach. Again, roughly similar prediction performances were achieved in a third study across several psychiatric hospitals comprising more than 40,000 inpatient episodes [29]. When training the model on all but one of the included institutions and predicting on the heldout hospital, the mean AUC for nonresponse was 0.65. Finally, a fourth study predicted remission in a sample of

approx. 2000 CBT outpatients who on the majority had a diagnosis of depression or anxiety and reported a balanced accuracy of 59% for the heldout set [30]. All of these studies used substantial samples sizes and powerful approaches for validation and additionally, all of these studies used sociodemographic and clinical data for building the predictive models, although the exact information included varied. Given these similarities, it is an interesting finding that prediction performances were in a comparable range. It remains to be seen whether such binary predictions can be made considerably more accurate without including new data sources.

A second trend is using the PAI [26] to predict which one of two available treatments will result in superior outcomes for a given patient. By identifying predictor variables that do and do not interact with the treatment, outcome predictions for the treatment a patient actually received and the treatment a patient did not receive can be made [26]. The difference between the predicted outcomes for the optimal (i.e., superior) and non-optimal (i.e., inferior) treatments is the PAI [26]. A PAI closely around zero indicates that both available treatments are predicted to perform comparably for a given patient, while PAIs with increasing absolute values indicate an increasing predicted benefit of one treatment over another [32]. The PAI can then be evaluated by comparing the outcome scores of patients who received their optimal treatment to those who did not; if the model is useful, patients who received their optimal treatment should have significantly better outcomes scores such as lower severity measures [26]. This approach has been used in a variety of studies to predict optimal treatment for depression, for example, for CBT versus antidepressants [26], CBT versus psychodynamic therapy [31], CBT versus IPT [32], or for comparing different CBT variants [33]. The PAI may be superior to binary (e.g., response versus non-response) outcome prediction as it uses a dimensional perspective and thus may also predict treatment superiority if outcomes for both treatments fall into the same overall outcome category. Additionally, while the specific statistical approach varied between some of the studies, the

overall PAI framework provides some similarity in the general approach. Intriguingly, statistical treatment outcome prediction using the PAI also outperformed predictions by clinicians in one study examining cognitive therapy versus IPT for depression [34]. Still, also for the PAI, specific predictive models from the individual studies have to be thoroughly evaluated in new, independent samples.

A third recurring theme is the predictive performance of anterior cingulate cortex (ACC) structure, function, and connectivity for treatment outcomes, particularly for depression and social anxiety disorder. For depression, a differential pattern of subcallosal ACC resting-state connectivity predicted remission after antidepressant and CBT treatment, respectively [35], and subgenual ACC activation predicted response to cognitive therapy in two independent samples [36]. In patients with social anxiety disorder, functional activation in the dorsal ACC during fMRI tasks predicted treatment response to Internet-delivered CBT with antidepressant or placebo [37] and, together with amygdala activation, to Internet-delivered CBT alone [38], while functional activation in the rostral ACC during emotion regulation predicted treatment response to face-to-face CBT [39]. Further results are reviewed elsewhere [40]. Despite these promising findings, it is important to note that a considerable variety of tasks, fMRI analytic methods, and machine learning approaches have been used in these studies. Furthermore, while prediction accuracies mostly ranged between 70% and 92%, sample sizes overwhelmingly were limited, risking overly optimistic estimates. Nevertheless, these studies provide a convincing foundation for future work examining the potential of the ACC for treatment outcome predictions.

Suicidality

Clearly, suicidality is one of the most severe symptoms associated with depression and, to a lesser degree, with anxiety disorders. Suicidality can be broken down in different stages with increasing proximity to actual suicide, from suicidal thoughts and ideation, to plans, and to suicide attempts. Particularly in moderate to severe

depression, suicidal thoughts are comparably prevalent, while plans and attempts are still less common. It is immediately clear that accurate prediction of emerging suicidality and particularly of suicide attempts would have tremendous clinical benefit and would provide a valuable tool to prevent actual suicides. Thus, it is a promising development that in recent years, a considerable number of studies aimed to use machine learning for individual prediction of suicidality.

In one exemplary study, suicide attempts were predicted for varying time intervals within the range of the next 7 days to the next 2 years from electronic health data using random forests [41]. The predictors from the electronic health data included for instance sociodemographics, diagnoses, medication data, past health care utilization, and past suicide attempts. One particular feature of this study is noteworthy: those without a future suicide attempt were selected from two groups: in one analysis, these individuals were hospital cases without a record of suicide attempts, while in the other analysis, these individuals also showed self-injury, but without suicidal intent. As expected, predicting future suicide attempts versus nonsuicidal self-injury was harder than predicting future suicide attempts versus general hospital cases, but both predictions showed very promising results, with AUCs of 0.84 and 0.92, respectively, for suicide attempts within the next 7 days. Also, prediction performance was only slightly decreasing for larger time intervals. Besides focusing on actual suicide attempts as a clinically relevant outcome, the study excels through its large sample size of approx. 5,000–15,000 patients for these analyses. An important point to keep in mind however is that the presented study requires that a suicidal individual has already established some sort of contact to the health system. As this cannot be taken for granted, there is increasing interest in using publicly available data for the prediction of suicidality (e.g., [42]).

Relapse

In contrast to other potential applications, the prediction of relapse, readmission, or rehospitalization versus lasting treatment success

has not received much attention so far, and only few studies are available. However, this might change, as a substantial proportion of patients experience reoccurrence of symptoms. Predictive models could be valuable to identify patients for receiving additional treatment doses such as booster sessions in psychotherapy or for intensified monitoring after treatment termination. Among the few studies available, one predicted relapse within 12 months after terminating CBT for depression and anxiety problems [43]. Four separate models were investigated, which used information gathered up to and including different time points (baseline only, treatment termination, 1-month follow-up, and 3-months follow-up). Relapse was predicted with an AUC of 0.72 based on baseline data and did only somewhat increase to a final AUC of 0.84 with more data becoming available, although prediction sensitivity and specificity varied considerably across the models. A second study is available that predicted rehospitalization within 2 years after treatment with most patients receiving medication [44]. Here, the best prediction achieved an AUC of 0.68. Although both studies are roughly comparable in their sample sizes of between 300 and 400 patients, a notable difference is that the first study only used sociodemographic and clinical predictors, while the second study also included structural MRI, blood serum, and sleep information. The benefit of this additional information was only moderate, however.

Outlook

Within the field of mental health, and particularly for depression and anxiety disorders, there is considerable interest in applying AI and machine learning methods under the broad framework of precision medicine. This approach bears great potential for improving the detection and treatment of mental disorders, but at the same time, there seems still a long way to go before clinical application. While sample sizes in some studies are becoming substantial and methods get more and more refined, most work is still on a “proof-of-concept” level. That means, initial results appear promising and worthy of further

investigation, but prediction accuracy may be far from clinical utility, or results are accompanied by considerable limitations. Moreover, predictive models are usually trained from zero for every study, but rarely applied across studies, which would be paramount to evaluate replicability and generalizability. Additionally, there may be considerable publication bias, with many prediction studies in small samples that do not provide significant results remaining unpublished. In order to move forward, more systematic research is needed that replicates and validates promising predictive models. In their excellent paper, Woo and colleagues give a formal characterization of the needed research program, where the most promising candidate models are examined further, from initial studies in single, small datasets to replication and generalization in other samples, sites, and settings until, eventually, the population level may be reached [45]. Across these stages, the predictive models also get more and more characterized regarding their performance overall and for specific groups and settings. Although originally aimed at neuroimaging models, this framework can be applied independently of data modality.

Considering such a research program emphasizes the need for datasets of substantial sample size. Ultimately, even representative samples for certain populations may be needed. Acquiring large datasets in the field of mental health and particularly for patient data is very challenging due to a number of reasons, including data privacy and confidentiality. The need for large sample sizes is furthermore related to the open question which kind of data is needed for good predictive models: it is probably not by chance that the studies with substantial sample sizes presented in the section on treatment outcome all used sociodemographic and clinical data for prediction, which is often either routinely collected in naturalistic settings or represents the backbone of data collection in large clinical trials. However, it is unclear whether predictive models suitable for clinical use can be constructed based on this data alone. It may be that more specific clinical assessments, for instance including etiologic or maintenance factors, or additional data modalities such as EMA,

MRI, or others are needed. In a very promising development, the last years have seen increasing efforts across many fields to acquire new large-scale datasets or to share and pool already collected data. Some examples in the neurosciences and in mental health may be the UK Biobank sample, the ENIGMA consortium, or the German KODAP initiative. These and similar initiatives may play a crucial role in the development and validation of AI and machine-learning-driven predictive models for depression and anxiety in the future.

Beyond the more and more extensive evaluation of prediction performance, the framework by Woo and colleagues also specifies that the performance of predictive models needs to be characterized more in depth with subsequent stages of testing [45]. At the moment, prediction performance is often given only by one or few metrics for the complete test sample. However, this may not be enough, given that it appears that statistical models may either discriminate against or produce incorrect results for certain groups, depending on the representation of these groups during the training stage of the model [46]. This may affect predictive models for depression and anxiety disorders as well. In the future, it will become more and more important to evaluate prediction performances in relationship to relevant sociodemographic and clinical characteristics (see recommendation in [47]). Again, this may speak in favor of collecting and using representative samples at some point. Finally, beyond the construction, optimization, and evaluation of the predictive models, there are a number of further ethical, legal, and societal aspects that may influence the development, application, and acceptance of predictive models in the field [48].

In conclusion, precision medicine bears great potential for improving the detection and treatment of depression and anxiety disorders, and AI and machine learning may be the most important tools for realizing this potential. However, while research interest increased considerably in recent years, the field is just starting to develop. Rigorous research with replication and evaluation of promising predictive models is needed in order to move toward clinical application.

Cross-References

- AIM and Explainable Methods in Medical Imaging and Diagnostics
- AIM and mHealth, Smartphones and Apps

References

1. GBD 2016 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet*. 2017;390(10100):1211–59.
2. Kessler RC, Petukhova M, Sampson NA, Zaslavsky AM, Wittchen HU. Twelve-month and lifetime prevalence and lifetime morbid risk of anxiety and mood disorders in the United States. *Int J Methods Psychiatr Res*. 2012;21(3):169–84.
3. Wittchen HU, Jacobi F, Rehm J, Gustavsson A, Svensson M, Jonsson B, et al. The size and burden of mental disorders and other disorders of the brain in Europe 2010. *Eur Neuropsychopharmacol*. 2011;21(9):655–79.
4. Gustavsson A, Svensson M, Jacobi F, Allgulander C, Alonso J, Beghi E, et al. Cost of disorders of the brain in Europe 2010. *Eur Neuropsychopharmacol*. 2011;21(10):718–79.
5. Chesney E, Goodwin GM, Fazel S. Risks of all-cause and suicide mortality in mental disorders: a meta-review. *World Psychiatry*. 2014;13(2):153–60.
6. Santoft F, Axelsson E, Ost LG, Hedman-Lagerlof M, Fust J, Hedman-Lagerlof E. Cognitive behaviour therapy for depression in primary care: systematic review and meta-analysis. *Psychol Med*. 2019;49(8):1266–74.
7. Khan A, Mar KF, Faucett J, Schilling SK, Brown WA. Has the rising placebo response impacted antidepressant clinical trial outcome? Data from the US Food and Drug Administration 1987–2013. *World Psychiatry*. 2017;16:181–92.
8. Loerinc AG, Meuret AE, Twohig MP, Rosenfield D, Bluett EJ, Craske MG. Response rates for CBT for anxiety disorders: need for standardized criteria. *Clin Psychol Rev*. 2015;42:72–82.
9. Ozmaro U, Wahlestedt C, Nemeroff CB. Personalized medicine in psychiatry: problems and promises. *BMC Med*. 2013;11:132.
10. van Loo HM, Cai T, Gruber MJ, Li J, de Jonge P, Petukhova M, et al. Major depressive disorder subtypes to predict long-term course. *Depress Anxiety*. 2014;31(9):765–77.
11. Drysdale AT, Gosenick L, Downar J, Dunlop K, Mansouri F, Meng Y, et al. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nat Med*. 2017;23(1):28–38.
12. Wang PS, Berglund P, Olfson M, Pincus HA, Wells KB, Kessler RC. Failure and delay in initial treatment contact after first onset of mental disorders in the

- National Comorbidity Survey Replication. *Arch Gen Psychiatry*. 2005;62(6):603–13.
13. Rosellini AJ, Liu S, Anderson GN, Sbi S, Tung ES, Knyazhanskaya E. Developing algorithms to predict adult onset internalizing disorders: an ensemble learning approach. *J Psychiatr Res*. 2020;121:189–96.
 14. Na K-S, Cho S-E, Geem ZW, Kim Y-K. Predicting future onset of depression among community dwelling adults in the Republic of Korea using a machine learning algorithm. *Neurosci Lett*. 2020;721:134804.
 15. Cohen JR, Thakur H, Young JF, Hankin BL. The development and validation of an algorithm to predict future depression onset in unselected youth. *Psychol Med*. 2020;50(15):2548–56.
 16. Zhang Y, Wang S, Hermann A, Joly R, Pathak J. Development and validation of a machine learning algorithm for predicting the risk of postpartum depression among pregnant women. *J Affect Disord*. 2021;279:1–8.
 17. Su D, Zhang X, He K, Chen Y. Use of machine learning approach to predict depression in the elderly in China: a longitudinal study. *J Affect Disord*. 2021;282:289–98.
 18. Wolfers T, Buitelaar JK, Beckmann CF, Franke B, Marquand AF. From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neurosci Biobehav Rev*. 2015;57:328–49.
 19. Varoquaux G. Cross-validation failure: small sample sizes lead to large error bars. *Neuroimage*. 2018;180 (Pt A):68–77.
 20. Boeke EA, Holmes AJ, Phelps EA. Toward robust anxiety biomarkers: a machine learning approach in a large-scale sample. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2020;5(8):799–807.
 21. Grotgerd D, Suslow T, Bauer J, Ohrmann P, Arrolt V, Stuhrmann A, et al. Discriminating unipolar and bipolar depression by means of fMRI and pattern classification: a pilot study. *Eur Arch Psychiatry Clin Neurosci*. 2013;263(2):119–31.
 22. MacMaster FP, Carrey N, Langevin LM, Jaworska N, Crawford S. Disorder-specific volumetric brain difference in adolescent major depressive disorder and bipolar depression. *Brain Imaging Behav*. 2014;8(1):119–27.
 23. Takizawa R, Fukuda M, Kawasaki S, Kasai K, Mimura M, Pu S, et al. Neuroimaging-aided differential diagnosis of the depressive state. *NeuroImage*. 2014;85(Pt 1):498–507.
 24. Pradier MF, Hughes MC, McCoy TH Jr, Barroilhet SA, Doshi-Velez F, Perlis RH. Predicting change in diagnosis from major depression to bipolar disorder after antidepressant initiation. *Neuropsychopharmacology*. 2021;46(2):455–61.
 25. Hilbert K, Lueken U, Muehlhan M, Beesdo-Baum K. Separating generalized anxiety disorder from major depression using clinical, hormonal, and structural MRI data: a multimodal machine learning study. *Brain Behavior*. 2017;7(3):e00633.
 26. DeRubeis RJ, Cohen ZD, Forand NR, Fournier JC, Gelfand LA, Lorenzo-Luaces L. The personalized advantage index: translating research on prediction into individualized treatment recommendations. A demonstration. *PLoS One*. 2014;9(1):e83875.
 27. Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatr*. 2016;3(3):243–50.
 28. Pradier MF, McCoy TH Jr, Hughes M, Perlis RH, Doshi-Velez F. Predicting treatment dropout after antidepressant initiation. *Transl Psychiatry*. 2020;10(1):60.
 29. Wolff J, Gary A, Jung D, Normann C, Kaier K, Binder H, et al. Predicting patient outcomes in psychiatric hospitals with routine data: a machine learning approach. *BMC Med Inform Decis Mak*. 2020;20(1):21.
 30. Hilbert K, Kunas SL, Lueken U, Kathmann N, Fydrich T, Fehm L. Predicting cognitive behavioral therapy outcome in the outpatient sector based on clinical routine data: a machine learning approach. *Behav Res Ther*. 2020;124:103530.
 31. Cohen ZD, Kim TT, Van HL, Dekker JJM, Driessen E. A demonstration of a multi-method variable selection approach for treatment selection: recommending cognitive-behavioral versus psychodynamic therapy for mild to moderate adult depression. *Psychother Res*. 2020;30(2):137–50.
 32. Huibers MJ, Cohen ZD, Lemmens LH, Arntz A, Peeters FP, Cuijpers P, et al. Predicting optimal outcomes in cognitive therapy or interpersonal psychotherapy for depressed individuals using the personalized advantage index approach. *PLoS One*. 2015;10(11):e0140771.
 33. Friedl N, Berger T, Krieger T, Caspar F, Grosse HM. Using the personalized advantage index for individual treatment allocation to cognitive behavioral therapy (CBT) or a CBT with integrated exposure and emotion-focused elements (CBT-EE). *Psychother Res*. 2020;30(6):763–75.
 34. van Bronswijk SC, Lemmens L, Huibers MJH, Peeters F. Selecting the optimal treatment for a depressed individual: clinical judgment or statistical prediction? *J Affect Disord*. 2021;279:149–57.
 35. Dunlop BW, Rajendra JK, Craighead WE, Kelley ME, McGrath CL, Choi KS, et al. Functional connectivity of the subcallosal cingulate cortex and differential outcomes to treatment with cognitive-behavioral therapy or antidepressant medication for major depressive disorder. *Am J Psychiatr*. 2017;174(6):533–45.
 36. Siegle GJ, Thompson WK, Collier A, Berman SR, Feldmiller J, Thase ME, et al. Toward clinically useful neuroimaging in depression treatment: prognostic utility of subgenual cingulate activity for determining depression outcome in cognitive therapy across studies, scanners, and patient characteristics. *Arch Gen Psychiat*. 2012;69(9):913–24.
 37. Frick A, Engman J, Alaie I, Bjorkstrand J, Gingnell M, Larsson EM, et al. Neuroimaging, genetic, clinical, and demographic predictors of treatment response in patients with social anxiety disorder. *J Affect Disord*. 2020;261:230–7.
 38. Mansson KN, Frick A, Boraxbekk CJ, Marquand AF, Williams SC, Carlbring P, et al. Predicting long-term outcome of internet-delivered cognitive behavior therapy

- for social anxiety disorder using fMRI and support vector machine learning. *Transl Psychiatry.* 2015;5:1–7.
39. Klumpp H, Fitzgerald JM, Kinney KL, Kennedy AE, Shankman SA, Langenecker SA, et al. Predicting cognitive behavioral therapy response in social anxiety disorder with anterior cingulate cortex and amygdala during emotion regulation. *Neuroimage Clin.* 2017;15:25–34.
40. Lueken U, Zierhut KC, Hahn T, Straube B, Kircher T, Reif A, et al. Neurobiological markers predicting treatment response in anxiety disorders: a systematic review and implications for clinical application. *Neurosci Biobehav Rev.* 2016;66:143–62.
41. Walsh CG, Ribeiro JD, Franklin JC. Predicting risk of suicide attempts over time through machine learning. *Clin Psychol Sci.* 2017;5(3):457–69.
42. O’Dea B, Wan S, Batterham PJ, Calear AL, Paris C, Christensen H. Detecting suicidality on twitter. *Internet Interv.* 2015;2:183–8.
43. Lorimer B, Delgadillo J, Kellett S, Lawrence J. Dynamic prediction and identification of cases at risk of relapse following completion of low-intensity cognitive behavioural therapy. *Psychother Res.* 2021;31(1):19–32.
44. Cearns M, Opel N, Clark S, Kaehler C, Thalamuthu A, Heindel W, et al. Predicting rehospitalization within 2 years of initial patient admission for a major depressive episode: a multimodal machine learning approach. *Transl Psychiatry.* 2019;9:285–94.
45. Woo CW, Chang LJ, Lindquist MA, Wager TD. Building better biomarkers: brain models in translational neuroimaging. *Nat Neurosci.* 2017;20(3):365–77.
46. Zou J, Schiebinger L. AI can be sexist and racist – it’s time to make it fair. *Nature.* 2018;559:324–6.
47. Cearns M, Hahn T, Baune BT. Recommendations and future directions for supervised machine learning in psychiatry. *Transl Psychiatry.* 2019;9(1):271.
48. Singh I, Rose N. Biomarkers in psychiatry. *Nature.* 2009;460(7252):202–7.



Artificial Intelligence for Autism Spectrum Disorders

113

Elisa Ferrari

Contents

Introduction	1580
AI Applications for ASD: Objectives, Data, and Challenges	1580
Objective-Based Categorization	1581
Field-Based Categorization	1583
Challenges of the AI-Based Research on ASD	1589
Conclusions	1590
Cross-References	1591
References	1591

Abstract

Autism spectrum disorder (ASD) is a chronic and extremely heterogeneous neurodevelopmental disorder, difficult to diagnose and with a still unclear multifactorial etiology. Given the scarce knowledge on this condition the research cannot be hypothesis-driven and must range across various sub-fields of biology and medicine, analyzing the big data produced by last generation healthcare and smart technologies. Artificial intelligence (AI) can

represent a valuable tool in this context, thanks to its ability to automatically discover complex patterns in high-dimensional data. This work represents a guide to the use of AI for research on ASD, describing various possible applications, which can differ for their objective (improving diagnosis, ranking severity, defining subtypes of ASD, drug discovery, etc.) and field (genetics, structural and functional neuro-imaging, etc.). For each application, the nature of the data and the most appropriate AI techniques to analyze them are described, along with illustrative examples of successful studies. The guide also includes a discussion on the major challenges currently affecting AI-based research on ASD: the lack of data and the subsequent problems of overfitting and confounding effects. Finally promising future research avenues, such as the adoption of explainable AI and ensemble learning are outlined.

Supplementary Information: The online version of this chapter (https://doi.org/10.1007/978-3-030-64573-1_249) contains supplementary material, which is available to authorized users.

E. Ferrari (✉)
Scuola Normale Superiore, Pisa, Italy
e-mail: elisa.ferrari@sns.it

Keywords

Autism spectrum disorder · Artificial intelligence · Machine learning · Deep learning · Neural networks · Neuroimaging · Magnetic resonance imaging · Electroencephalography · Genetics · Confounders

Introduction

Autism spectrum disorder (ASD) is a chronic and extremely heterogeneous neurodevelopmental disorder characterized by deficits in communication and social interaction, in addition to restricted and repetitive patterns of behavior [1]. According to the Autism and Developmental Disabilities Monitoring Network, the latest estimate of ASD prevalence in the USA is 1 in 59 [2], with a male to female ratio of approximately 3:1. Symptoms may be visible already at 6 months of age, becoming explicit around the second or third year, and may vary widely along lifespan.

Several evidences suggest a genetic liability for ASD. For instance, the individual risk of ASD is much higher in families with one member already diagnosed with it: approximately 10 times greater in subjects with a sibling with ASD [3]. Furthermore, in these families even undiagnosed members can show autism-related traits, termed broader autism phenotype. Despite substantial progress has been made in understanding ASD genetic underpinnings, most ASD-associated DNA variations account for no more than 1% of ASD cases and very few of them lead to ASD in every carrier [4]. All these considerations suggest a *complex or multi-factorial* nature of ASD [5]. According to various researchers in fact the ASD etiology may be explained in terms of a multifactorial threshold model, meaning that ASD is caused by the accumulation of both genetic and environmental risk factors over a certain threshold that determines a diagnosable phenotype. In such a model, the broader phenotype can be explained by the presence of a lower, but still high, number of risk factors (see Fig. 1).

The scarce knowledge on the causes of ASD limits the progress in the research and development of targeted therapies; however, many psychological treatments exist, that can improve the quality of life of the subjects with ASD. Such treatments proved to be more effective when started at an early age and thus an early diagnosis is of fundamental importance [6]. However, diagnosing ASD is a difficult task requiring psychiatrist assessment, since no quantitative biomarkers exist and the heterogeneous behavioral symptoms can be easily misunderstood or ignored during infancy.

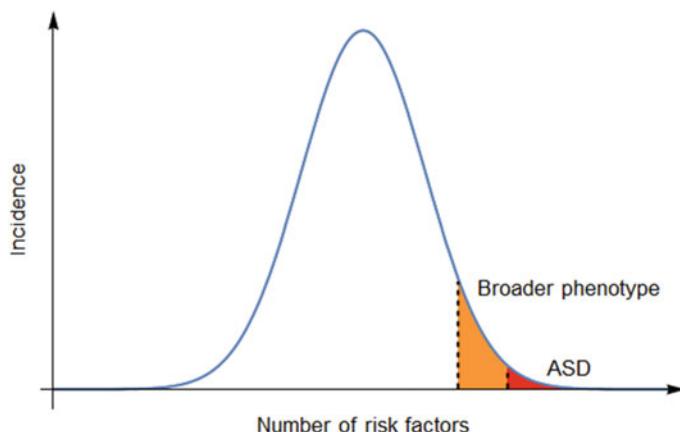
Because of the multifactorial nature of ASD and the lack of knowledge about it, the research on this disorder cannot be hypothesis-driven and must range across various sub-fields of biology and medicine, analyzing high-dimensional data such as DNA sequence, gene expression, and neuroimaging data. In this kind of open ended research, Artificial Intelligence (AI) can be very promising, because of its ability to autonomously discover complex patterns in large datasets, requiring minimal preprocessing and data domain expertise.

AI Applications for ASD: Objectives, Data, and Challenges

As detailed in the introduction, the scarce knowledge on ASD and its heterogeneous etiology and phenotype allow a wide variety of approaches to study this condition. In this section some of the most important and promising AI applications to study ASD are presented and categorized according to two criteria: the objective of the analysis and the research field. Challenges in the application of AI methods are also discussed at the end of this section. A brief presentation of all these contents is also provided in (Video 1).

Throughout the text, several AI methods are simply mentioned while for some a brief and qualitative description is reported. In any case an in-depth discussion of the algorithms is out of the scope of this chapter. More details about clustering, conventional machine learning (ML) (such as logistic regression, support vector machine,

Fig. 1 Multifactorial threshold model applied to ASD. According to this model, ASD is caused by the accumulation of certain amount of risk factors, which are naturally present in the population, but are more numerous in ASD subjects. The broader phenotype can be explained by the presence of a lower but still high number of risk factors



random forest, etc.) and multivariate regression-based methods can be found in this book [7], while a comprehensive reference for deep learning (DL) method is provided in [8].

Objective-Based Categorization

In this section, different approaches to study ASD are outlined, subdivided by their objective. Although several different aims are described, by far the most common one is *automatic/early diagnosis*. In fact, as previously said, ASD diagnosis is currently very complex and somewhat subjective, and thus may greatly benefit from the discovery of an objective classifier based on a quantitative biomarker.

Automatic/Early Diagnosis

Automatic diagnosis is the typical application for which a categorical classifier based on ML or DL techniques is employed. Basically, a classification algorithm is trained to distinguish between two classes of subjects, the ASDs and Healthy Controls (HCs), which are represented by a set of data, that can range from medical images and records to domestic videos of the subject behavior. The objective is to obtain an accurate and sensitive predictor able to provide a diagnosis on new subjects with acceptable performance. If a reliable classifier is found, it can be used for computer-aided diagnosis or for a screening program to identify subjects at risk. If this algorithm also

performs well on subjects of pre-verbal age, it can be used for *early diagnosis*. This application may have a large social impact, given that treatments started before the first year of life can significantly improve the quality of life of ASD subjects.

A wide range of studies encompassing several fields of research can be categorized under this label and many examples will be provided in section “[Field-Based Categorization](#).” Particularly encouraging results have been obtained on electroencephalography (EEG) data with standard ML methods, and in video analysis with more complex DL approaches. However, these fields are less explored, compared for example to magnetic resonance imaging (MRI), and the few studies available have been conducted on small datasets, making the assessment of their reliability more difficult.

Severity Recognition

Severity recognition is mainly performed using a multivariate regression approach. Usually a ML algorithm is trained to reproduce a score measuring symptom severity or disability degree of input subjects. Typical scores reconstructed can be ASD-specific such as ADOS and ADI [9], but also more generic indicators such as Intelligence Quotient (IQ) or Verbal Intelligence Quotient. As for *automatic/early diagnosis* applications, the input data representing each subject can be very heterogeneous, but the most common approach consists in trying to infer an ASD-related score from brain MRI data. The objective of this kind of

studies is to understand if the severity perceived at the behavioral level and measured with such scores can be extracted from biomedical data, meaning that there is a measurable biological feature that correlates with such scores. An example of a *severity recognition* study is described in [10]. The authors used a multivariate regression method on structural MRI (sMRI) data to predict the ADOS score (ranging from 0 to 10) achieving an average absolute error of 1.36. Similar results have been obtained also using task-based functional MRI (fMRI) data [11].

Subtypes Definition

Considering the heterogeneous nature of ASD, another interesting approach consists in finding more homogeneous subgroups of ASD using clustering algorithms. Decomposing the heterogeneity might help in the definition of ASD subtypes that individually may be easier to distinguish from HCs. Furthermore, given that ASD heterogeneity ranges across multiple sectors (i.e., genetics, neural systems, cognition, behavior, development, and clinical topics), a possible approach consists in comparing data-driven subtypes defined in different fields. An example of this technique is given by [12], in which severity-based subtypes of ASD obtained with clustering methods applied to ADI and ADOS information have been genetically analyzed. An increased genotype similarity within subjects belonging to the same subtype was observed, suggesting that subjects with similar symptom severity may have similar genetic etiology.

Longitudinal Studies

Given that ASD symptoms can naturally change during lifespan or in response to treatments and stressful/traumatic events, longitudinal studies can be very important to understand the disease. However, collecting data of the same subjects along time can be very resource-intensive and this limits the number of works of this kind. Despite these studies are mainly designed with the same aim and methodology of the previous applications, they deserve a distinct category because the data collection process and the conclusions drawn are substantially different.

For instance, a *longitudinal study* for *early diagnosis* may use data collected multiple times during early childhood to train an ML to distinguish ASDs from HCs analyzing the temporal evolution of the subject records. This approach has been adopted using multiple behavioral and developmental measures acquired between 8 and 36 months of age, showing that already at 14 months ASD and broader atypical development can be classified with an area under the receiver operating characteristic curve (AUC) of 71% [13]. Another possible approach, tailored for *severity recognition*, consists in using past records to predict disease evolution after a certain temporal interval. With this aim, using a multivariate regression on features extracted from resting-state fMRI (rs-fMRI) data, it has been shown that it is possible to predict improvements in adaptive behavior scores after more than 1 year with 100% sensitivity and 71% precision [14].

Explorative Analysis

The approaches belonging to the first three categories described above can be evaluated with quantitative measures, such as accuracy, sensitivity, and AUC for categorical classifications, mean squared error for regressions, overlap, and centroid distance for clusterings. However, as it will be better explained in section “[Challenges of the AI-Based Research on ASD](#),” in ASD research it is very difficult to achieve good and reproducible results. For this reason, a valuable study must not be restricted to the sole production of a satisfactory result as measured by the aforementioned metrics, but must also include an analysis of the data that mostly drove such result. For example, an ML/DL-based analysis should include an explanation of the decisional process learned from the data; a regression-based one should provide a measure of the variance explained by each independent variable; a clustering-based one should report the features that characterize each cluster.

Thus, in this objective oriented categorization, the distinct label of *explorative analysis* is dedicated to studies that include the mentioned explanations. Providing an intelligible interpretation of the results also facilitates the comparison with other studies, which would be otherwise

impracticable using only performance-based metrics, and can provide new insights about the disease even without achieving state-of-the-art performance.

For instance, in one recent *explorative analysis* [15] for *automatic diagnosis* based on sMRI derived features and reaching an AUC of 79%, the authors performed a wide permutation test to isolate the most reliable part of the entire decisional pattern found by the classifier. Surprisingly, they showed that this reduced pattern is still able to capture mean differences between ASD and HC classes, using substantially different data, that is, acquired with different modalities and recruitment criteria. This study proves the importance of explainability in ML/DL applications in order to collect reliable results.

Drug Discovery

Given that the exact causes and a common biomarker of ASD are still unknown, it is very difficult to conduct drug discovery studies and the few ones currently ongoing are not based on AI, but on more conservative approaches. However, as already said AI is perfectly suited for this kind of open ended analysis involving large number of data. For instance, it has been shown that a DL model trained with experimental data of drug effects on singular cells can predict pharmacological effects (such as toxicity, efficacy, therapeutic use) on the entire human organism [16]. In the specific case of ASD, after the discovery of some protein-coding genes directly involved in syndromic forms of the disorder, the use of AI has been suggested for *target druggability* applications, that is, identification of the genes, out of the ones involved in the disease, that are likely to be successfully targeted by existing drugs [17].

Teaching and Interaction

Children with ASD have from moderate to severe difficulties in socialization, verbal and non-verbal communication, and understanding other perspectives: abilities that significantly affect the learning process. For this reason, computer-aided learning has been extensively analyzed proving that ASD children show a preference for electronic screen media, computer-generated speech, and

game-like elements. In this context, AI can really make the difference, allowing targeted learning and personalized interaction. For instance, a smartglasses-based augmented reality system based on AI has been developed to teach ASD children life skills that may facilitate or enhance self-sufficiency. The device implements a DL algorithm that recognizes facial expressions viewed by the user and teaches to him/her how to interpret them [18]. Another interesting application is the development of robots able to interact with humans and the environment, thanks to the continuous AI-based analysis of audio and video data collected in real-time. This application seems very promising, in fact, it appears that ASD subjects pay more attention and reduce the repetitive behaviors when interacting with the robots rather than with people [19].

Field-Based Categorization

This section provides a summary of the main fields of research on ASD where AI can make the difference. Each field description focuses on the peculiarities of its data and on which AI techniques are most suited to them. To show the potential impact of each field, some illustrative studies are reported. Since, as explained in the previous section, ASD research is commonly oriented toward *automatic/early diagnosis*, most of the studies reported are based on the use of ML/DL techniques to classify ASDs from HCs; however, all the other approaches detailed in the objective-based categorization are also possible in the fields mentioned here.

Brain Structural Imaging

Brain structural imaging can be done essentially with computed tomography (CT) and sMRI. However, the last one is usually preferred for research purposes because it allows a better diversification of the brain soft tissues and it does not use ionizing radiations, thus it is totally safe also for children.

SMRIs are 3D images, depending on the spatial resolution adopted, they can be composed by up to tens of millions of voxels (i.e., 3D pixels),

and, when decompressed, they usually occupy tens of MBs.

Due to their large volume, typically these images are processed using dedicated algorithms to extract meaningful properties and features, such as volume, surface area, thickness of various brain regions, or structural connectivity via diffusion tensor imaging techniques. Then conventional ML methods are usually applied to these features for *automatic/early diagnosis* applications. Alternatively, raw or minimally pre-processed sMRIs can be analyzed using DL algorithms which accomplish the feature extraction process and classification integrally. The most commonly used DL algorithms are convolutional neural networks (CNNs) or their derivatives, which proved to be highly performant in extracting relevant features while taking into consideration spatial dependencies important in images [8] (see Fig. 2).

Thanks to its non-invasivity and high spatial resolution, MRI (both functional and structural) is the most used technique to study ASD and to date there is a high number of papers on the application of AI to MRIs, mainly for *early/automatic diagnosis* and for the discovery of a quantitative biomarker of the disorder. Studies have been conducted both on small cohorts of homogeneously scanned subjects and on large, multicenter datasets of images acquired using different modalities, scanners, and recruitment criteria. Despite this large body of literature, a consensual anatomical biomarker for ASD based on sMRI is still missing. In fact, the results obtained are very

heterogeneous and often not significant, since they allow to distinguish ASDs from HCs with an accuracy often lower than 60% [20] on the largest datasets, while the few higher-performing results are rarely replicated in subsequent studies or with bigger samples [21]. However, particularly worth mentioning is a prospective study [22] conducted on a large sMRI dataset of infants between 6 and 12 months with both high and low risk for developing the disorder, aimed at understanding a long debated observation: early increased brain volume in ASD development. The authors showed that hyperexpansion of the cortical surface area precedes brain overgrowth observed in 15 high-risk infants who were subsequently diagnosed with ASD. Relying on this assumption, they trained a DL algorithm with structural features extracted from MRIs encoding the growth of brain regions, achieving a true positive rate of 81% and a sensitivity of 88% in predicting ASD in high-risk children [22].

Brain Functional Imaging

Brain functional imaging can be done with different techniques, each capturing different properties of brain activity. The most employed ones are: positron emission tomography (PET), EEG, and fMRI.

PET provides brain metabolic information. During a PET exam, a radioactive tracer bound to a molecule of metabolic interest is injected into the patient and, thanks to its emissions, the scanner can reconstruct the spatial distribution of the molecule, revealing in which tissues it is

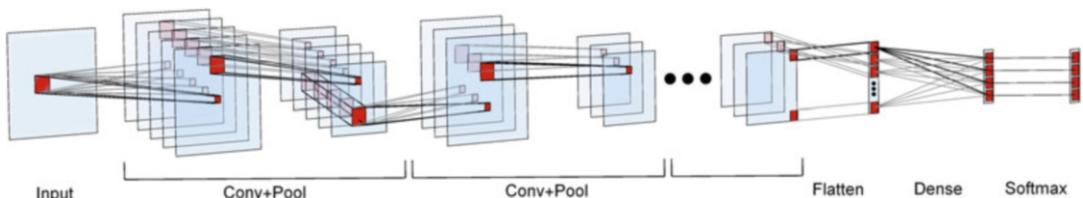


Fig. 2 Structure of a CNN. A series of convolutional and pooling layers is followed by a flattening operation and a sequence of dense layers, ending in the classification layer (Softmax), which contains the probability of the input of belonging to each possible class. A convolutional layer extracts spatial (or temporal) features through stackable receptive fields (called kernels) with tunable size. The

size of the kernel determines the resolution with which the layer sees the input, while the depth of the layer determines the level of abstraction of the feature extracted. More shallow layers will therefore recognize lines, points, and basic shapes, while deeper layers will extract more complex and complete features, up to the characteristics that allow the final classification

accumulated. With PET it is possible to acquire a single image, that recapitulates the distribution of the molecule after a certain period of assessment, or a video recorded while the subject is performing a certain task, to analyze the metabolic changes during that task. Usually PET images are composed by between 500,000 and 5,000,000 voxels, depending on the spatial resolution adopted and videos usually acquire 1–10 frames per minute. The dimension of a PET scan can be roughly estimated in the range of tens of MBs per frame, as an sMRI.

Given that PET requires exposition to ionizing radiation and an intravenous line for radioactive tracer injection, it is not suited for research purposes on young subjects. However, for its ability to investigate complex metabolic process in the brain, not otherwise analyzable, PET-based research has been carried on involving small groups of adults. The scarcity of participants currently limits the possibility to apply AI approaches, but the important perspectives of this field and the need to analyze high-dimensional and complex data suggest that PET-based studies will be a promising application for AI.

For ASD research, the study of cerebral blood flow (CBF) and glucose metabolism (GM) along with the most peculiar PET application, that is, pharmacokinetics, are particularly promising. Traditional data analysis studies on CBF consensually show the presence of hypo-perfusion of temporal lobes and in speech related areas in ASD subjects, while with GM an atypical metabolism in cingulate, parietal, and occipital cortex during memory tasks is observable in ASD subjects. A rare example of the application of AI on PET data consists in the use of standard ML algorithms for *automatic diagnosis* to discover multivariate combination of hypo- and hyper-perfusion characterizing ASD brain. The classifier trained reached an accuracy of 88% and an in-depth analysis of the decisional pattern shows that ASDs present a hypo-perfusion in right superior temporal sulcus and hyper-perfusion in the contralateral postcentral area of the brain [23].

Electroencephalography (EEG) records the electrical activity of the brain, by detecting the

electrical pulses that propagate along neuron sequences aligned with pairs of electrodes placed on the head. The main limitation of EEG is the difficulty to place the electrodes in the same exact position for each patient, due to the high variability of head dimensions. This clearly represents an important obstacle to data comparison across a cohort of subjects and to studies aggregating data from multiple sites.

EEG data consist in a set of time series, called channels, that depending on the protocol adopted can vary between 3 and 130. Single raw channels are acquired with a sampling frequency between 100 and 500 Hz. Thus, for task-based studies, with acquisition duration in the order of the tens of minutes, the data size can easily exceed 50 MB. EEG data are very difficult to interpret in a medical perspective, thus usually they are analyzed extracting statistical features typically employed to describe time series, such as average, standard deviation, signal power (i.e., the average of the squares of its values), and their corresponding measures in the Fourier space. All these features (and possibly others) are calculated on different frequency bands, known to contain electrical signals from brain activity and conventionally named with Greek letters.

According to a recent review on this matter, many studies have been conducted using EEG-based features to train ML algorithms for ASD diagnosis, reaching very high accuracy values [24]. In fact, the reference reports 10 recent studies obtaining between 90% and 100% of accuracy. To date, this application is one of the most promising for finding an ASD quantitative biomarker, but it should be considered that available studies are usually conducted on small cohorts (usually less than 40 subjects). Furthermore as it has been extensively described for MRI applications, it is likely that also for EEG the equipment and the infrastructure used to record data (i.e., electrode locations, number of channels, sampling rate) may strongly influence the trained classification algorithm, making classifying EEG data acquired with a different setup almost impossible. Another issue is the difficulty in giving a physiological interpretation of a possible EEG-based biomarker for ASD, which would

corroborate the findings. However, in the specific case of a syndromic form of ASD caused by the duplication of a portion of the 15th chromosome (and thus called Dup15q), an EEG biomarker with a reasonable physiological interpretation has been found [25]. In fact, it has been observed that subject with Dup15q presents higher signals in the beta frequencies band with respect to both non-syndromic ASDs and HCs. This signature similarly appears in subjects treated with drugs that enhance the effect of the GABA neurotransmitter, which is also regulated by genes that are duplicated in the subjects with Dup15q.

Functional MRI (fMRI) allows to detect the changes in the magnetic properties of the brain due to the transport of oxygenated blood in correspondence of neuronal activity, providing thus an indirect measure of spatial and temporal brain activity. FMRI exams can be done while the subject is performing a certain task, to analyze which brain areas are involved in the task, or at rest to study spontaneous brain activity (resting state fMRI, rs-fMRI). This last exam is much simpler to perform on ASD subjects which, depending on the severity of the disease, may not be very collaborative in task execution.

FMRIs are 4D images (three spatial dimensions and a temporal one); depending on the resolutions adopted and on the duration of the task for fMRI or on the protocol followed for rs-fMRI, an acquisition can contain between 10 and 100 millions of data points, corresponding to 20–200 MB.

Similarly to sMRIs, fMRIs are usually pre-processed using dedicated toolboxes and pipelines which collect multiple important steps, such as skull removal, temporal filtering, head motion correction, and slice timing correction, that is, the temporal alignment of signals coming from different slices of the brain which have been acquired asynchronously. Another common step consists in reducing the number of temporal signals, which can be done condensing the signals of the voxels belonging to the same brain regions, or with a completely data-driven approach performing an independent component analysis. These reduced signals are then fed into a standard ML or transformed in correlation matrices (known

as functional connectivity matrices) and analyzed as images with DL. Alternatively, the raw data, minimally preprocessed, are used to train a 4D CNN, which treats the temporal dimension exactly as the spatial ones, or a recurrent neural network (RNN), specifically designed for temporal analysis [8]. The characteristic of the RNNs that makes them particularly suited for temporal analysis is that they process the temporal signals as streams of data and they contain memory retaining layers that simultaneously process the current input and the previous ones (see Fig. 3).

The previous discussion on the scarce reproducibility of sMRI research also applies to fMRI and rs-fMRI analyses. One of the best results in classification obtained to date in a large multi-site dataset reaches only an AUC of 70% [26]. The approach used was quite standard: the rs-fMRI data were reduced to functional correlation matrices and then fed into a DL algorithm trained to distinguish ASDs and HCs. Until now, the adoption of more sophisticated approaches such as the use of an RNN directly on time series has not improved this result, reaching an AUC of 68% [27].

Genetics

Genetic data comprise various kinds of data (i.e., genome, transcriptome, proteome, and metabolome), which can be analyzed in different ways.

The genome sequence is the complete list of the three billion nucleotide pairs composing the DNA, but usually only two types of genetic variations are stored and analyzed: the single nucleotide polymorphisms (SNPs, genetic variations relative to a reference genome involving a single nucleotide), which are roughly four to five million in a person's genome, and the copy number variations (CNVs, gain or loss of segments of genomic DNA relative to a reference) which are above 70 per genome with a mean size of 10^5 nucleotides.

In literature, some SNPs and CNVs that individually may be implicated in ASD etiology have been found. For instance, strong evidences are related to the SNP rs2217262 in the DOCK4 gene, which increases risk of ASD between two to four times, and the CNV known as Dup15q

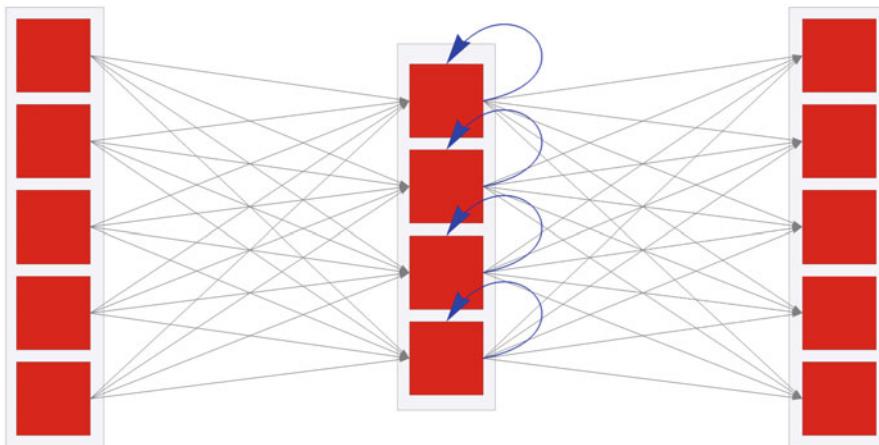


Fig. 3 Structure of an RNN. An RNN is based on the concept of memory. Each layer retains memory of what has previously passed through it, thanks to connections to itself (the blue arrows in the picture). These connections allow

the analysis and extraction of features from data that are not simultaneously available to the layer (such as for example in a video stream, or in any variable-length input)

which causes the homonym syndromic form of ASD. However, being ASD a complex disease, the multivariate approach typical of AI tools can help to unravel the articulated interactions of combinations of genetic variations, understanding how genetic risk factors confer susceptibility to ASD. A study shows that using SNP data it is possible to develop a regression based classifier correctly predicting ASD diagnosis in up to 80% of the cases in two datasets composed mainly by European subjects, however losing accuracy in a genetically dissimilar cohort of Asiatic people [28]. Similarly, another work [29] shows that a Random Forest-based classifier trained with CNV data can explain 7.9% of ASD while misclassifying only 3% of HCs, which is an important result considering that to date only 10% of ASD are expected to be caused by CNVs.

Another genome-level type of data, rarely analyzed but very promising for *early diagnosis* applications, is DNA methylation, an epigenetic modification by which methyl groups can be added to cytosine or adenine bases (about 4% of the cytosines are methylated in the human genome) changing in this way the activity of a DNA segment without changing its sequence. In a recent study it has been shown that a simple logistic regression using DNA methylation data can predict ASD with an AUC of 95% in newborns [30].

Transcriptome, proteome, and metabolome describe respectively which and how many gene transcripts, proteins, and metabolites are present in a cell or in a group of cells of a specific tissue. Thus, all these data can be represented as a vector of weights, of a size up to roughly 240,000 different transcripts, 80,000 proteins, and 120,000 metabolites; however, these numbers can significantly vary depending on the reference library used and to which extent these elements are grouped in similar functional sets. Being tissue-specific measures, transcriptome, proteome, and metabolome of brain tissues would be of particular interest in the study of ASD. Unfortunately, this kind of analysis is possible only on post-mortem samples, which limits the number of available data and as a consequence the possibility to analyze them with AI techniques. However, research on data from other tissues still yielded good results: in a classification study based on blood transcriptome, an AUC of 70% has been reached [31], while on plasma metabolome an accuracy of 81% has been obtained [32].

Besides the typical ML approaches for *automatic/early diagnosis* that have been mentioned so far, all the genetic data are particularly apt to be analyzed with clustering techniques, especially with hierarchical clustering algorithms, in order to explore the genetic heterogeneity of the disease

and to define genetically-grounded subtypes of ASD. For instance, a typical analysis pipeline of transcriptome includes the construction of a gene co-expression network, that is, a network showing closer connections between the highly correlated genes, and then the application of hierarchical clustering algorithms to define gene modules with coordinated biological functions. With this approach, one of the largest studies on brain post-mortem transcriptomic data showed that synaptic and glial functions along with inflammatory pathways are important brain downstream mechanisms characterizing ASD [33].

Finally, another important and promising field for exploring the genetic bases of ASD is risk gene discovery with ML or DL methods. Basically, the idea consists in moving from the traditional analyses to discover new genes implicated in the disease based on the comparison of HCs and ASDs genetic data, to a machine learning approach using genome-scale data (such as prior scores of genetic association, brain gene expression, and topological information from large gene interaction networks) as predictors to identify new genes with similar properties to already established ASD risk genes [34].

Video and Sensor Analysis

Neuroimaging and genetic data can be obtained only in medical structures, thus, when the subject has been already diagnosed or has a familial risk to develop ASD. Promising applications for *early detection*, likely life-impacting in the near future, are based on the study of data that can be acquired at home and analyzed remotely via telemedicine, such as home-made videos and sensor data that can be obtained with mobile or wearable devices. A high number of studies tried to automatically detect ASD through the analysis of children video, reaching high classification performances using either 3D CNNs (two spatial directions and one temporal), RNNs, or other more advanced network architectures derived from the previous ones. However, most of the classifiers developed expect input videos to portray the subject while performing specific actions. One recent study instead shows that it is possible to build a classifier able to recognize ASD typical repetitive behaviors

from arbitrary-length home-made videos of the children, without requiring them to do specific activities [35]. The classifier built reached an accuracy of 95%. From a methodological point of view, this work built a DL model made of four components, a 3D CNN used to extract spatiotemporal features followed by a temporal pyramid network (TPN), which progressively identifies longer and more abstract types of actions. In the third component the intermediate representations of this TPN are used to identify single actions typical of ASD children. Finally, the fourth component uses the TPN top-layer to predict the presence of repetitive behaviors. Interestingly, the presence of the third component forces the entire network to focus exclusively on ASD behaviors, such as hand flapping, head banging, spinning, toe walking, and moving fingers.

This example shows that in applications for which a large body of prior knowledge is available, such as the visual recognition of ASD typical behaviors, it is possible to fruitfully adopt more complex AI architectures exploiting highly annotated datasets, thanks to an implicit knowledge transfer from humans to machines. On physiological data, such as EEG and MRI, it is instead more difficult to guide the AI, which is in fact used to gain new insights on the disease adopting often simpler and more interpretable models.

Besides video analysis, there are a lot of other applications that study established behavioral characteristics of ASD, such as the sensor-based analyses of gait patterns, eye-movement, and repetitive gestures. In fact, using DL methods based on CNNs, accuracies up to 96% have been reported on this kind of data [36]. It must be noted that a long-standing and consensual result observed in this kind of studies is that ASD subjects present an eye-attention map focused toward objects rather than people, as it happens instead in HCs [37].

Despite all the applications described in this section reach very high performance in detecting ASD and are promising for remote and early diagnosis, due to the nature of the data, they cannot provide new insights neither on the causes nor on the possible treatments of the disorder.

Miscellaneous

The heterogeneity of ASD manifestations and the still very uncertain causes of the disorder keep the way open for an unlimited number of studies, which for the sake of synthesis cannot be fully reported in this chapter. For instance, given that retinal changes have been found in ASD, an AI-based classifier for retinal images was developed showing a promising AUC of 97%, although being tested on a small cohort of subjects [38]. Similarly, the higher incidence and severity of gastrointestinal problems in ASDs with respect to HCs and the plurireported alterations of the gut microbiome in the disorder, suggested the development of a machine learning diagnosing tool based on intestinal microbiome, which achieved an AUC of 83% [39]. These few examples show how many fields and types of data can be involved in the research on ASD. In this context, AI is a versatile and promising tool, not only because it allows to examine complex data structures such as the ones described in this chapter but also because it can be used to combine multi-modal data without making any prior assumption on them. This has already been done for data that are somehow related, such as sMRI and fMRI [40], but considering the heterogeneity of ASD, it might be interesting to combine even very different modalities, possibly using ensemble methods, that is, machine learning techniques that combine several classifiers in order to produce a single optimal predictive model.

Challenges of the AI-Based Research on ASD

One of the most appetible and legendary feature of AI is its ability to automatically discover patterns from raw data, that is, without any preprocessing step. This power has a price: the access to a high number of examples, which is a long-standing issue in medical applications in general and for research on ASD in particular. The following problems obstruct the data collection process:

- The heterogeneity of ASD, which can be represented only collecting data from a large

and diversified sample, possibly including healthy relatives and subjects belonging to the broader phenotype

- The necessity to collect data during early infancy, with all the issues related: privacy, family consensus, and collaboration from the children during data collection/acquisition
- The difficulty in the diagnosis, which may delay the acquisition of data

The scarcity of data, affecting all the fields of research on ASD, is exacerbated by their high dimensionality, which causes problems of overfitting and leads to wrong results driven by spurious correlations. This phenomenon may explain why in MRI research a significant drop in classification performance has been reported going from the initial studies based on small cohort of patients to larger ones. Whether the good results obtained in less investigated fields with limited access to data may suffer from the same overfitting problem is still not fully understood and deserves further investigations.

In addition, another important factor that may mislead an AI-based study is the presence of the so called confounders, that is, variables affecting the data that do not represent clinically relevant aspects, but might nevertheless bias the analysis if they are unequally represented in the dataset. In fact, biomedical data, but also video and sensor data, depend on a high number of variables: demographics information, acquisition/recording modalities and instruction given to the patient, personal physiological characteristics, medications, and feature assessment criteria (such as qualitatively defined scores or indices); thus it is common that one or more of such variables are biased in the dataset under examination. Confounder effects are well known in various medical applications, but in a recent work it has been shown that their effect is particularly severe in MRI-based ASD research using ML/DL algorithms [15]. The authors used the Confounding Index (CI) [41], an index measuring how much a confounder variable can affect a classification problem, to quantify and compare the effects of commonly cited sources of heterogeneity in sMRI-based research. The results (Fig. 4) show

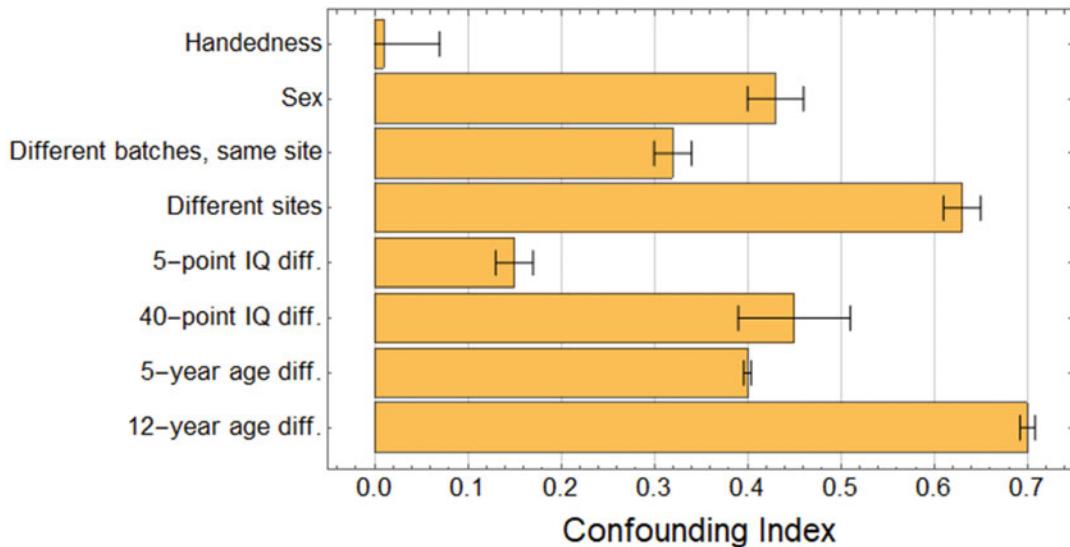


Fig. 4 CI value of several possible confounders for ASD. The CI is an index that ranks from 0 to 1 the effect of a confounder variable on a certain binary classification task.

For ASD, sex, acquisition modalities, acquisition site, large IQ differences, and even small age differences are all strongly confounding. (Data from Ref. [15])

that there are a large number of factors that have a strong confounding effect on the ASD/HC classification task, including nonbiological ones such as acquisition site. Given that the analysis was conducted on highly processed data, the unexpected strength of the site effect is supposedly enhanced by two phenomena: the complexity of the classification task caused by the heterogeneity of the disorder and the ability of AI to find unrealistically complex multivariate patterns. As a proof for this hypothesis, the authors trained the same algorithm several times to classify subjects scanned in different sites, using a variable number of randomly selected highly processed features. As it can be noted from Fig. 5, the algorithm needs at least 20 features (regardless of which ones) to understand the site provenience, proving that individually the features are substantially independent from the acquisition site, but ML can discover subtle and highly multivariate dependencies which are difficult to pinpoint (and thus correct). This ability that might be considered a strength of AI methods turns out to potentially drive misleading effects. Considering that the difficulty of the ASD classification task and that the use of ML/DL approaches is common to several important

research works on ASD, this problem may not be circumscribed solely to MRI data.

Summarizing, besides the collection of larger datasets, which might prevent overfitting, another important challenge for AI-based research is the adoption of confounding-resilient methods, which are currently object of study in applied computer science.

Conclusions

Despite a high number of studies ranging different domains have been conducted, current understanding of ASD is still incomplete and research has a long way to go. AI is a powerful and versatile methodology to address open-ended research involving high-dimensional data, when prior knowledge necessary for hypothesis-driven analyses is missing. Currently, the most common application is the adoption of ML/DL approaches to discover quantitative biomarkers to improve the diagnosis process and allow early treatment intervention. Across the various fields that can be explored in this way, MRI is the most studied one, because it represents a safe technology that can

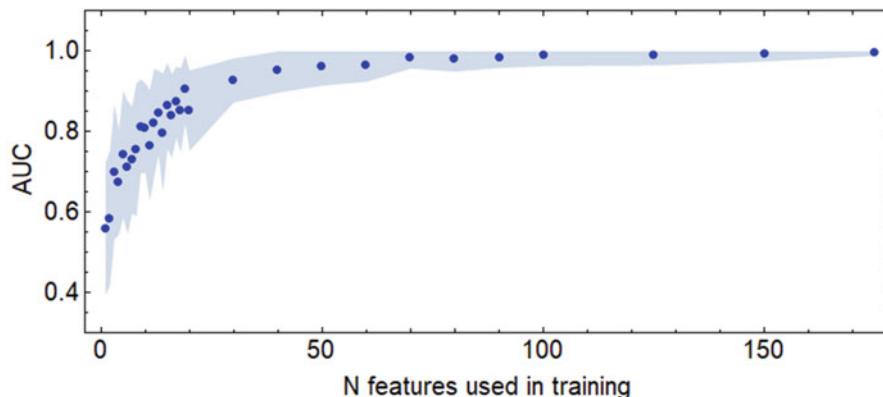


Fig. 5 AUC obtained vs number of features used in the training process of a classifier distinguishing MRIs of subjects scanned in two different sites. The classifier needs at least about 20 features to reach good performance. The

relatively tight error band in light blue, representing different instances of training with a different set of features, shows that the results do not significantly depend on the specific features chosen. (Data from Ref. [15])

investigate ASD brain structural and functional abnormalities with high spatial resolution, and thus is potentially useful both to find a biomarker and to understand the neurological causes of the disorder. However, the large body of literature produced contains highly heterogeneous results, often difficult to reproduce and sometimes even contradictory. The lack of a consensual result has been imputed mainly to the diffuse presence of overfitting and confounding effects, which can be highly misleading in data-driven approaches.

The next major challenges for AI-based research on ASD are thus the collection and release of large datasets and the development of new AI techniques to deal with confounding effects. Furthermore, the adoption of explainability approaches, which allow a better understanding of the decisional process made by the algorithm, might help to spot and correct for the mentioned problems and at the same time ease the discovery of new insights on the disorder.

In conclusion, given the heterogeneity of ASD, research should move forward on multiple fields, possibly integrating multimodal data, encouraging the adoption of explainable AI and ensemble methods (which may capture the multi-faceted nature of ASD) and keeping up with the development of new AI-based tools to deal with overfitting and confounding effects.

Cross-References

- ▶ [AIM and Explainable Methods in Medical Imaging and Diagnostics](#)
- ▶ [AIM in Clinical Neurophysiology and Electroencephalography \(EEG\)](#)
- ▶ [AIM in Pharmacology and Drug Discovery](#)
- ▶ [Applying Principles from Medicine Back to Artificial Intelligence](#)

References

1. American Psychiatric Association, et al. Diagnostic and statistical manual of mental disorders (DSM-5®). American Psychiatric Pub, Arlington, VA; 2013.
2. Baio J, Wiggins L, Christensen DL, Maenner MJ, Daniels J, Warren Z, Kurzus-Spencer M, Zahorodny W, Rosenberg CR, White T, et al. Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, united states, 2014. MMWR Surveill Summ. 2018;67(6):1.
3. Sandin S, Lichtenstein P, Kuja-Halkola R, Larsson H, Hultman CM, Reichenberg A. The familial risk of autism. JAMA. 2014;311(17):1770–7.
4. Fernandez BA, Scherer SW. Syndromic autism spectrum disorders: moving from a clinically defined to a molecularly defined approach. Dialogues Clin Neurosci. 2017;19(4):353.
5. Ivanov HY, Stoyanova VK, Popov NT, Vachev TI. Autism spectrum disorder—a complex genetic disorder. Folia Med. 2015;57(1):19–28.

6. Koegel LK, Koegel RL, Ashbaugh K, Bradshaw J. The importance of early identification and intervention for children with or at risk for autism spectrum disorders. *Int J Speech-Lang Pathol.* 2014;16(1):50–6.
7. Bonacorso G. Machine learning algorithms. Packt Publishing Ltd, Birmingham, UK; 2017.
8. Goodfellow I, Bengio Y, Courville A. Deep learning. MIT Press; 2016. <http://www.deeplearningbook.org>.
9. Le Couteur A, Haden G, Hammal D, McConachie H. Diagnosing autism spectrum disorders in pre-school children using two standardised assessment instruments: the adi-r and the ados. *J Autism Dev Disord.* 2008;38(2):362–72.
10. Moradi E, Khundrakpam B, Lewis JD, Evans AC, Tohka J. Predicting symptom severity in autism spectrum disorder based on cortical thickness measures in agglomerative data. *NeuroImage.* 2017;144: 128–41.
11. Zhuang J, Dvornek NC, Li X, Ventola P, Duncan JS. Prediction of severity and treatment outcome for ASD from fMRI. In: Rekik I, Unal G, Adeli E, Park SH, editors. Predictive intelligence in medicine. Cham: Springer International Publishing; 2018. p. 9–17.
12. Veatch OJ, Veenstra-VanderWeele J, Potter M, Pericak-Vance MA, Haines JL. Genetically meaningful phenotypic subgroups in autism spectrum disorders. *Genes Brain Behav.* 2014;13(3):276–85.
13. Bussu G, Jones EJH, Charman T, Johnson MH, Buitelaar JK, BASIS Team, et al. Prediction of autism at 3 years from behavioural and developmental measures in high-risk infants: a longitudinal cross-domain classifier analysis. *J Autism Dev Disord.* 2018;48(7): 2418–33.
14. Plitt M, Barnes KA, Wallace GL, Kenworthy L, Martin A. Resting-state functional connectivity predicts longitudinal change in autistic traits and adaptive functioning in autism. *Proc Natl Acad Sci.* 2015;112(48): E6699–706.
15. Ferrari E, Bosco P, Calderoni S, Oliva P, Palumbo L, Spera G. Maria Evelina Fantacci, and Alessandra Retico. Dealing with confounders and outliers in classification medical studies: the autism spectrum disorders case study. *Artif Intell Med.* 2020;108:101926.
16. Aliper A, Plis S, Artemov A, Ulloa A, Mamoshina P, Zhavoronkov A. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Mol Pharm.* 2016;13(7):2524–30.
17. Uddin M, Wang Y, Woodbury-Smith M. Artificial intelligence for precision medicine in neurodevelopmental disorders. *npj Digital Med.* 2019;2(1):1–10.
18. Liu R, Salisbury JP, Vahabzadeh A, Sahin NT. Feasibility of an autism-focused augmented reality smartglasses system for social communication and behavioral coaching. *Front Pediatr.* 2017;5:145.
19. Costa AP, Charpiot L, Lera FR, Ziafati P, Nazarikhoram A, Van Der Torre L, Steffgen G. More attention and less repetitive and stereotyped behaviors using a robot with children with autism. In: 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). IEEE, Manhattan, New York, U.S.; 2018. p. 534–9.
20. Haar S, Berman S, Behrmann M, Dinstein I. Anatomical abnormalities in autism? *Cereb Cortex.* 2014;26(4):1440–52.
21. Abraham A. Learning functional brain atlases modeling inter-subject variability. PhD thesis, Université Paris-Saclay, 2015.
22. Hazlett HC, Gu H, Munsell BC, Kim SH, Styner M, Wolff JJ, Elison JT, Swanson MR, Zhu H, Botteron KN, et al. Early brain development in infants at high risk for autism spectrum disorder. *Nature.* 2017;542 (7641):348–51.
23. Duchesnay E, Cachia A, Boddaert N, Chabane N, Mangin J-F, Martinot J-L, Brunelle F, Zilbovicius M. Feature selection and classification of imbalanced datasets: application to pet images of children with autistic spectrum disorders. *NeuroImage.* 2011;57(3): 1003–14.
24. Brihadiswaran G, Haputhanthri D, Gunathilaka S, Meedeniya D, Jayaratna S. EEG-based processing and classification methodologies for autism spectrum disorder: a review. *J Comput Sci.* 2019;15(8):1161–1183.
25. Frohlich J, Senturk D, Saravanapandian V, Golshani P, Reiter LT, Sankar R, Thibert RL, DiStefano C, Huberty S, Cook EH, et al. A quantitative electrophysiological biomarker of duplication 15q11.2-q13.1 syndrome. *PLoS One.* 2016;11(12):e0167179.
26. Heinsfeld AS, Franco AR, Craddock RC, Buchweitz A, Meneguzzi F. Identification of autism spectrum disorder using deep learning and the abide dataset. *NeuroImage Clin.* 2018;17:16–23.
27. Dvornek NC, Ventola P, Pelpfrey KA, Duncan JS. Identifying autism from resting-state fMRI using long short-term memory networks. In: International workshop on machine learning in medical imaging. Springer, Berlin, DE; 2017. p. 362–70.
28. Skafidas E, Testa R, Zantomio D, Chana G, Everall IP, Pantelis C. Predicting the diagnosis of autism spectrum disorder using gene pathway analysis. *Mol Psychiatry.* 2014;19(4):504–10.
29. Engchuan W, Dhindsa K, Lionel AC, Scherer SW, Chan JH, Merico D. Performance of case-control rare copy number variation annotation in classification of autism. *BMC Med Genet.* 2015;8(S1):S7.
30. Bahado-Singh RO, Vishweswariah S, Aydas B, Mishra NK, Yilmaz A, Guda C, Radhakrishna U. Artificial intelligence analysis of newborn leucocyte epigenomic markers for the prediction of autism. *Brain Res.* 2019;1724:146457.
31. Kong SW, Collins CD, Shimizu-Motohashi Y, Holm IA, Campbell MG, Lee I-H, Brewster SJ, Hanson E, Harris HK, Lowe KR, et al. Characteristics and predictive value of blood transcriptome signature in males with autism spectrum disorders. *PLoS One.* 2012;7 (12):e49475.
32. West PR, Amaral DG, Bais P, Smith AM, Egnash LA, Ross ME, Palmer JA, Fontaine BR, Conard KR, Corbett BA, et al. Metabolomics as a tool for discovery

- of biomarkers of autism spectrum disorder in the blood plasma of children. *PLoS One.* 2014;9(11):e112445.
33. Parikhshah NN, Swarup V, Belgard TG, Irimia M, Ramaswami G, Gandal MJ, Hartl C, Leppa V, de la Torre Ubista L, Huang J, et al. Genome-wide changes in lncrna, splicing, and regional gene expression patterns in autism. *Nature.* 2016;540(7633):423–7.
34. Brueggeman L, Koomar T, Michaelson JJ. Forecasting risk gene discovery in autism with machine learning and genome-scale data. *Sci Rep.* 2020;10(1):1–11.
35. Tian Y, Min X, Zhai G, Gao Z. Video-based early asd detection via temporal pyramid networks. In: 2019 IEEE International Conference on Multimedia and Expo (ICME). IEEE, Manhattan, New York, U.S.; 2019. p. 272–7.
36. Jaiswal S, Valstar MF, Gillott A, Daley D. Automatic detection of ADHD and ASD from expressive behaviour in RGBD data. In: 2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017). IEEE, Manhattan, New York, U.S.; 2017. p. 762–9.
37. Boraston Z, Blakemore S-J. The application of eye-tracking technology in the study of autism. *J Physiol.* 2007;581(3):893–8.
38. Lai M, Lee J, Chiu S, Charm J, So WY, Yuen FP, Kwok C, Tsui J, Lin Y, Zee B. A machine learning approach for retinal images analysis as an objective screening method for children with autism spectrum disorder. *EClinicalMedicine.* 2020;28:100588.
39. Wu T, Wang H, Lu W, Zhai Q, Zhang Q, Yuan W, Zhennan G, Zhao J, Zhang H, Chen W. Potential of gut microbiome for detection of autism spectrum disorder. *Microb Pathog.* 2020;149:104568.
40. Sen B, Borle NC, Greiner R, Brown MRG. A general prediction model for the detection of ADHD and autism using structural and functional MRI. *PLoS One.* 2018;13(4):e0194856.
41. Ferrari E, Retico A, Bacciu D. Measuring the effects of confounders in medical supervised classification problems: the confounding index (ci). *Artif Intell Med.* 2020;103:101804.



Artificial Intelligence in Schizophrenia

114

Howard Schneider

Contents

Introduction	1596
Artificial Intelligence Applied to the Research, Diagnosis, and Treatment of Schizophrenia and Related Disorders: Pre-2000	1597
Artificial Intelligence Applied to the Research, Diagnosis, and Treatment of Schizophrenia and Related Disorders: 2000–2012	1598
Artificial Intelligence Applied to the Research, Diagnosis, and Treatment of Schizophrenia and Related Disorders: 2012–2018	1599
Artificial Intelligence Applied to the Research, Diagnosis, and Treatment of Schizophrenia and Related Disorders: 2019–Present	1601
Current and Future Clinical Use of AI Techniques in the Diagnosis and Treatment of Schizophrenia and Related Disorders	1604
Cross-References	1605
References	1605

Abstract

Principles and methods in artificial intelligence applied to the research, diagnosis, and treatment of schizophrenia and related disorders are reviewed from the 1980s to 2020. Support vector machines (SVMs), neural networks,

expert systems, deep learning neural networks, autoencoders, deep belief networks, random forests, ensemble methods, and cognitive architectures are some of the AI techniques reviewed as applied to schizophrenia and related disorders. Connectionist models are used to better explain the development of schizophrenia. SVMs and deep learning neural networks are used to help interpret neuroimaging data in patients with schizophrenia. Deep learning networks are used to predict various outcomes for patients with schizophrenia. Deep learning neural networks and SVMs are used in the schizophrenia drug discovery process. AI-powered avatar therapy and socially

Supplementary Information: The online version of this chapter (https://doi.org/10.1007/978-3-030-64573-1_214) contains supplementary material, which is available to authorized users.

H. Schneider (✉)
Sheppard Clinic North, Toronto, ON, Canada
e-mail: hschneidermd@alum.mit.edu

interactive robots are used in the therapy of patients with schizophrenia. At the time of this writing, use of AI techniques is starting to switch from the research setting to the clinical setting to help diagnose and treat patients with schizophrenia or schizophrenia-related disorders. For example, a multimodal machine model combined with human evaluations can fairly accurately predict the transition to psychosis for high-risk young individuals.

Keywords

Schizophrenia · Psychosis · Artificial intelligence · Machine learning · Precision psychiatry · Precision medicine · Personalized medicine · Subsymbolic · Symbolic · Avatar therapy

Introduction

Artificial intelligence (AI) is defined by Russell and Norvig [1] in terms of inclusion as a member in one of four broad categories – thinking rationally, acting rationally, thinking humanly, or acting humanly. In the “thinking rationally” class, an AI system applies logical reasoning to facts in order to solve a problem. In the “acting rationally” class, the AI system acts somewhat autonomously to obtain the best possible outcome of some goal given the existing constraints. In the “thinking humanly” class, the AI system should think like a human, thus there must essentially be cognitive modeling on the part of the AI system in this category of the definition. In the “acting humanly” class, an AI system successfully approaching the popular Turing Test (i.e., the AI system functions so well that a human asking written questions in a different room, could not tell whether the responses are from another human or from an AI system) is required. Here the AI system needs to have the following abilities – natural language processing (to communicate with the questioner), knowledge representation, automated reasoning, and machine learning (to improve knowledge representation as well as to recognize patterns). Applications of artificial intelligence to areas of

schizophrenia fall in all these areas, as well as in areas that are more statistical in nature and less in keeping with the categories above (Video 1).

The Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5), classifies schizophrenia within the chapter on “Schizophrenia Spectrum and Other Psychotic Disorders” [2]. Schizophrenia diagnostic criteria require that there must be, for much of the time, at least one month of delusions, hallucinations, or incoherent or derailed speech, resulting in a reduced level of functioning. As well there must be at least two of the following: delusions, hallucinations, disorganized speech, catatonic or very disorganized behavior, and/or negative symptoms such as reduced emotions. While as noted above the active symptoms must be present for at least a month, there must also be active or residual symptoms for 6 months. The diagnosis requires exclusion of affective disorders with psychotic features, autism spectrum disorder (although comorbid diagnoses are possible), and physiological or medical causes of symptoms.

In the chapter of the DSM-5 on “Schizophrenia Spectrum and Other Psychotic Disorders,” in addition to schizophrenia, there are other related disorders listed [2]. These disorders include delusional disorder, brief psychotic disorder, schizophreniform disorder, schizoaffective disorder, substance/medication-induced psychotic disorder, psychotic disorder due to another medical condition, catatonia associated with another mental disorder, and schizotypal personality disorder, although in the latter details are found in the DSM-5 chapter on “Personality Disorders.” While below the focus is principally on applications of AI in the etiology, diagnosis, and treatments in schizophrenia, there will be consideration, where relevant, of AI applications toward any of these DSM-5 schizophrenia-related disorders.

In the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10), in the chapter on “Mental and Behavioural Disorders” there is a block on “Schizophrenia, Schizotypal and Delusional Disorders” [3]. This group includes: schizophrenia, schizophreniform disorder/psychosis,

schizophrenia unspecified, schizotypal disorder, persistent delusional disorders, acute and transient psychotic disorders, induced delusional disorder, schizoaffective disorders, other nonorganic psychotic disorders, and unspecified nonorganic psychosis. Although the focus below is mainly on the application of AI for schizophrenia, applications of AI for the study, diagnosis, or treatment of any of these ICD-10-described schizophrenia-related disorders will be included where relevant.

The field of artificial intelligence is often arbitrarily divided into what are termed “Symbolic AI” and “Subsymbolic AI” approaches. Symbolic AI uses symbolic formulations of problems and symbolic logical solutions, i.e., there is manipulation of discrete symbols. Symbolic AI is sometimes referred to as “good old-fashioned AI (GOFAI)” due to the more current predominance of subsymbolic AI approaches. In subsymbolic AI there is not a formulation of the problem or its solution with a particularly human-obvious symbolic-like representation. For example, in artificial neural networks (ANNs), typically referred to as “neural networks,” there is not an obvious symbolic representation of a problem, but instead there is a massive collection of nodes (or “artificial neurons”) with connections (or “synapses”) to each other in a variety of wiring arrangements. Below both symbolic and subsymbolic approaches to AI applied to schizophrenia and related disorders will be included.

Artificial Intelligence Applied to the Research, Diagnosis, and Treatment of Schizophrenia and Related Disorders: Pre-2000

Artificial intelligence techniques started to become more consistently applied to the field of schizophrenia in the mid-1980s. Work by Hoffman used specialized Hopfield neural networks to simulate models of psychosis induction [4]. However, during this era, in general in the field of AI, expert systems were the most predominantly used paradigm for AI systems, and started to find applications in psychiatry. An expert system is a

symbolic approach to AI, utilizing a large collection of if-then rules obtained from a human expert, and software that applies these if-then rules to the properties of a problem that is to be solved. Work by Maurer and colleagues in the late 1980s compared an expert system that diagnosed schizophrenia from patient data with diagnoses obtained via the more traditional classification systems, with the expert system giving similar results [5].

By the 1990s expert systems in many different areas of AI had not performed as well as had been hoped for, and the entire AI field received less funding and attention, often referred to as the “AI Winter.” Artificial neural networks (or “neural networks”), essentially related to the modern deep learning artificial neural networks that would later be developed, were at this time unfortunately not thought to hold much promise. However, work did slowly continue on artificial neural networks. Cohen and Schreiber in 1992 described a neural network model which explored attention and language deficits in schizophrenia [6], and Hoffman and McGlashan in 1993 used artificial neural networks to model a breakdown in corticocortical communication [7].

Symbolic reasoning algorithms had long been part of the AI field, and during this time, indeed continued to be used in applications related to schizophrenia. For example, work by Garfield and Rapp in 1994 applied semantic networks with node and pathway-based reasoning rules to psychotic speech [8]. A semantic network is essentially a diagram representing knowledge, with lines between the nodes representing concepts, i.e., a symbolic approach to AI. In Garfield and Rapp’s work, reasoning rules operated on the nodes of and the lines in the semantic network diagram, and could recognize what they termed “crazy talk” from the violations of the reasoning rules.

As noted by Seeman, work on artificial neural network models of schizophrenia started to increase modestly by the mid-1990s, with the hope that these models could provide better explanations for the pathophysiology in schizophrenia [9]. Work by Lowell and Davis used artificial neural networks (ANNs) for the more practical

purpose of predicting the length of hospital stay based on schizophrenic and other psychiatric patient data the ANNs were trained on [10]. Work by Hoffman and McGlashan in 1998 used ANNs to show that decreased cortical connectivity modeled the development of auditory hallucinations [11]. Work by Corson and colleagues in 1999 used an ANN to identify and measure the caudate nucleus in neuroimages of first-episode psychosis patients and control subjects [12].

Artificial Intelligence Applied to the Research, Diagnosis, and Treatment of Schizophrenia and Related Disorders: 2000–2012

During this era there was work on both symbolic and subsymbolic artificial intelligence approaches to schizophrenia and related disorders. For example, Razzouk and colleagues in 2006 described using decision support systems, which are essentially expert systems, to help the practicing physician in the clinical diagnosis of schizophrenia [13].

As neural network theory and technology slowly improved over this decade, there was the obvious application of using machine learning to extract indicators of schizophrenia and related disorders from neuroimaging. For example, Jafri and Calhoun used ANNs to attempt to detect the presence of schizophrenia from brain fMRIs [14]. Bose and colleagues in 2008 describe using an ANN model to distinguish schizophrenic patients from controls in PET imaging with an 89% sensitivity and 94% specificity [15].

Support vector machines (SVMs) started to be more widely used during this time period to classify features in a variety of datasets obtained from research on patients with schizophrenia. Support vector machines are not actually machines but a type of supervised machine learning model, useful for classifying data into a particular category as well as finding relationships between variables. As shown below in Fig. 1, the SVM model attempts to find the best hyperplane between different classes of datapoints. Similar to a neural network, there is a training phase where annotated

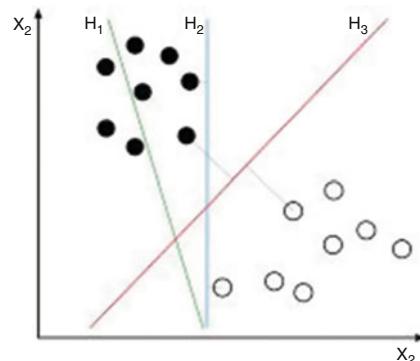


Fig. 1 SVM model creates a “hyperplane” that allows classification of data points. In this simple low-dimension example note that H₃ separates the classes with a better margin than H₁ or H₂. (Creative Commons License BY-SA. Credit to Zack Weinberg)

training examples (i.e., annotated with the correct category an example belongs to) allows the SVM to build a model of the data. The SVM can then be used to classify data it has not seen before. Modifications of the SVM algorithm allow it to also function in an unsupervised fashion and find clustering of similar items in a large collection of data. Struyf and colleagues in 2008 used a variety of classification algorithms including SVMs, nearest shrunken centroids, decision trees, naïve Bayes, and nearest neighbor classification to classify patient genomics, demographics, and clinical data as bipolar disorder versus schizophrenia [16]. SVMs classified significantly better than the other algorithms used. Ozyurt and Brown in 2009 described a variety of AI techniques including SVMs and probabilistic reasoning, for retrieving knowledge from the scientific literature including schizophrenia abstracts [17].

The nodes or “neurons” in an ANN, while originally inspired by biological neurons, are artificial constructs quite different and much simplified compared to a natural neuron, as is their organization or wiring. There is an interest in the computational neuroscience community for more biologically realistic neural networks. Unlike in ANNs, biological neurons are much more sparsely and recurrently connected to each other [18]. Given that a variety of hippocampal abnormalities are found in schizophrenia, more realistic hippocampal computer simulations, for example,

were thought to help better understand its etiology [19]. ANNs continued to be used to model aspects of the pathology of schizophrenia. For example, Karolidis and colleagues in 2010 used ANNs to model a number of cloned molecules which could bind to human dopamine D1 and D2 receptors [20].

Artificial Intelligence Applied to the Research, Diagnosis, and Treatment of Schizophrenia and Related Disorders: 2012–2018

The theory and technology behind various machine learning approaches, including ANNs with multiple hidden layers and termed “deep learning,” had started to greatly improve in the mid-2000s. In 2012 work by Krizhevsky, Sutskever, and Hinton using deep learning won a computer vision competition by a large margin over older methods [21]. This achievement is regarded as an approximate start of what is called the “deep learning revolution” and propelled the utilization of deep learning into many domains, including the research and clinical aspects of the field of schizophrenia. In the ImageNet contest a computer-based system needed to classify as accurately as possible large numbers of different images into some thousand different classes. Krizhevsky and colleagues used a deep convolutional neural network. In such a neural network there are multiple layers of nodes (or “neurons”) connected with layers where the nodes act as convolutional layers where such layers extract features from the previous layers preserving the spatial relationships but essentially mapping into a small-size receptive field and extracting features as such. Krizhevsky and colleagues used 650,000 neurons arranged in a number of layers including five convolutional layers. Within a few years deep learning neural networks improved (and grew in size and processing power) to the point where they could classify the ImageNet images more accurately than human competitors.

A review by Veronese and colleagues in 2013 gives an overview of machine learning approaches

in schizophrenia but describes little of deep learning [22]. However, within a few years deep learning was being used in the field. In 2016 Kim and colleagues note that deep neural networks (DNNs) with multiple hidden layers were performing much better in classification tasks compared to SVMs and earlier AI models. They used a DNN to obtain functional connectivity patterns from resting-state functional magnetic resonance imaging [23]. A review by Arbabshirani and colleagues in 2017 considers the application of machine learning techniques including the emergence of deep learning to the prediction of brain disorders from patient neuroimaging data [24].

From 2016 onwards there was increasing application in the research and clinical aspects of schizophrenia of not only deep learning but also a variety of artificial intelligence techniques. For example, innovative work by Miotto and colleagues used a deep learning network incorporating autoencoders on a hospital-wide patient database to extract in an unsupervised fashion actionable features on individual patients such as personalized prescription, disease prediction, and clinical trial recruitment with better predictions for patients with schizophrenia than many other disorders [25]. Miotto and colleagues used a stack of particular denoising autoencoder layers. An autoencoder is a neural network that has an encoder layer of neurons that as its name suggests encodes an input signal. The encoded signal, usually reduced in dimensionality, i.e., there is a reduction in size of the input via nonlinear modifications, is then reconstructed by the decoder layer of neurons to give an output that maps in some way to the input signal. Thus, the output is not an exact copy of the input but a transformation of the input signal and, depending on how the autoencoder is constructed as well as combined with multiple layers of other autoencoders, can allow unsupervised learning to produce an efficient coding of the input signal which can extract or classify features of the input data (Fig. 2).

An interesting area of robotics in the 2010s was the development of “socially interactive robots” which are engineered to improve human-robot interactions by displaying nonverbal cues including facial emotions. Since nonverbal social

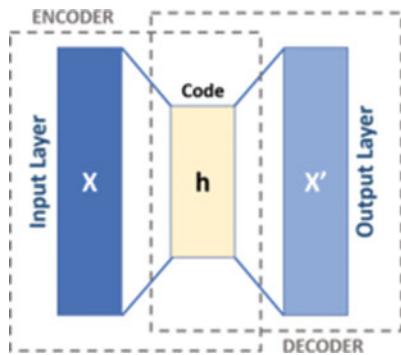


Fig. 2 Overview of an autoencoder. (Creative Commons License BY-SA. Credit to Michaela Massi)

interactions are known to be poorer in patients with schizophrenia, Raffard and colleagues in 2016 considered the use of a humanoid iCub robot, shown in Fig. 3 albeit without emotions activated in this photograph, in patients with schizophrenia [26]. It was found that both patients and controls recognized better the emotional facial expressions of humans than robots. Thus, the authors conclude that while humanoid robots have a theoretical potential in the role of increasing social functioning in patients, this study should only be considered exploratory.

Work by Arnon and colleagues in 2016 approached the theoretical realm of what is possible using modern electronics and AI technologies. Arnon and colleagues described the potential use of thought-controlled nanoscale robots which are activated by a magnetic field which is produced when an external sensor detected a particular EEG pattern, and can do a particular function such as make available bioactive molecules that could treat schizophrenia [27].

Use of SVM methods continued in the 2010s. For example, work by Mikolas and colleagues in 2016 used SVM analysis on the resting-state fMRI images of patients with a first-episode schizophrenia spectrum disorder and healthy controls, and was able to distinguish the anterior insula connectivity of patients from controls with an accuracy of 73.0% [28]. Work by Zarogianni and colleagues used SVM analysis in predicting schizophrenia based on a self-completed measure

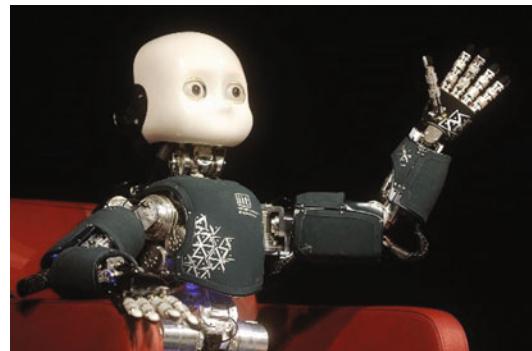


Fig. 3 iCub socially interactive robot, neutral emotions. (Creative Commons License BY-SA. Credit to Niccolò Caranti)

of schizotypy, a declarative memory test and a structural MRI brain scan [29].

A deep belief network is a type of deep neural network with a stack of layers of simpler often unsupervised learning neural models. Each layer in the deep belief network extracts a higher level of features of the input signal. Deep belief networks had been used for extracting information from and recognizing images, for example. Work by Pinaya and colleagues trained a deep belief network on features from MRI brain morphometry to distinguish between patients with schizophrenia and healthy controls with an accuracy of 73.6% [30].

A technique known as random forest creates many different decision trees during training and later will use the class decided by, for example, the most decision trees. Work by Pergola and colleagues in 2017 used random forests methodology to classify grey matter volume in the thalamus of MRIs from patients with schizophrenia, non-affected siblings, and controls, and found a familial relation associated with reduced thalamic grey matter volume in patients with schizophrenia [31].

Just as older AI techniques in the 1980s were used to model possible mechanisms in the development of psychosis, this continued in the 2010s with the now popular deep learning neural networks. For example, Keshavan and Sudarshan discuss how deep learning algorithms can possibly overemphasize objects

the network believes are recognized leading to outputs that are detached from reality, and thus may possibly model some mechanisms involved in psychosis [32].

As in other fields of medicine, mobile apps incorporating AI features began to emerge with increasing frequency, especially from 2016 onwards. Work by Bain and colleagues in 2017 describe a mobile app to assess medication adherence by visual confirmation for patients with schizophrenia [33]. Work by Birnbaum and colleagues in 2017 used machine learning of Twitter (American blogging service) feeds from users self-diagnosed with schizophrenia in an attempt to classify users of this blogging service with schizophrenia from users who were considered healthy controls [34].

Machine learning started to become commonplace in imaging. For example, Dluhoš and colleagues describe creating a meta-model by combining SVM classifiers trained on local datasets and combining data from multiple centers [35]. Honnorat and colleagues used a clustering machine learning method that indicated a distinction between schizophrenia patients and controls in the temporal-thalamic-peri-Sylvian regions, frontal regions, and thalamus [36].

Work by Winterburn and colleagues compared logistic regression, SVMs, and linear discriminant analysis on cortical thickness and tissue density estimates, in patients with schizophrenia and healthy controls, with most accuracies between 55% and 70% [37]. Zeng and colleagues in 2018 used a deep discriminant autoencoder network to learn and distinguish fMRI functional connectivity features from patients with schizophrenia from healthy controls [38]. Bae and colleagues used SVM analysis to show from a public fMRI dataset that there were significant differences in patients with schizophrenia from healthy controls, especially in the anterior right cingulate cortex, the superior right temporal region, and the inferior left parietal region [39]. Work by Mikolas and colleagues used SVM to distinguish diffusion tensor MRI data from patients with first-episode

schizophrenia spectrum disorder and healthy controls with an accuracy of 62.3% [40].

Artificial Intelligence Applied to the Research, Diagnosis, and Treatment of Schizophrenia and Related Disorders: 2019–Present

Toward the end of the 2010s a variety of AI techniques, including DNNs and SVMs, started to be applied to more relevant challenges in schizophrenia research, diagnosis, and treatment. AI methods allowed better interpretation of neuroimaging of patients with schizophrenia. Machine learning was being used in schizophrenia drug discovery research. AI techniques were being used in clinical schizophrenia diagnosis and care.

Zhao and So used DNNs and SVMs to predict the repositioning of schizophrenia drugs based on their drug expression profiles as candidates for other diseases [41]. Work by Lin and colleagues used three different machine learning algorithms (logistic regression, naïve Bayes classifier, and C4.5 decision tree) to examine two G72 SNP genotypes and G72 (D-amino-acid oxidase activator, DAOA) protein levels (previously associated with schizophrenia patients) to try to classify schizophrenia patients from healthy controls. The G72 protein levels alone gave most of the effect, with the naïve Bayes giving the best specificity of 0.95 and the logistic regression technique yielding the most sensitivity of 0.88 [42].

Fond and colleagues used machine learning via decision trees of data from patients with controlled schizophrenia to predict relapse at two years. High anger, high physical aggressiveness, high lifetime psychiatric hospitalizations, low education level, and high positive symptoms at baseline were the strongest predictors of relapse [43].

Work by Kalmady and colleagues examined resting-state fMRI data from patients with schizophrenia who had not been treated with antipsychotic medications and healthy controls. They

used an ensemble machine learning classifier that was able to classify a resting fMRI into the correct schizophrenia/healthy control class with an accuracy of 87% versus 53% accuracy expected by chance alone [44]. An ensemble method uses multiple machine learning algorithms to obtain better accuracy than any one of the individual machine learning algorithms that it is based on. Kalnady and colleagues called their ensemble model “EMPaSchiz” which stood for “Ensemble algorithm with Multiple Parcellations for Schizophrenia prediction.”

Work by Brodley and colleagues in 2019 trained SVMs from the responses of patients to the Early Psychosis Screener for Internet (EPSI) and predicted if the patient would be diagnosed with a psychotic disorder in 12 months [45]. Barrera and colleagues used digitally assisted nursing observations (i.e., obtained via computer vision and signal processing) in an acute mental health input ward with patients with schizophrenia, and found that patients’ safety was not compromised [46]. Wu and colleagues trained an ensemble machine learning method on the records of 70% patients with first-episode schizophrenia to predict a successful antipsychotic medication selection. Success was defined as not switching medications and not being hospitalized in the next 12 months. The remaining 30% patients’ data was not used for training, but kept for testing the machine learning model. If the individualized treatment which the machine learning method suggested was used, then the treatment success was 51.7% versus the actually observed 44.5% success [47].

Parola and colleagues used Bayesian classifier networks and found that pragmatic linguistic (language as well as other expressive means) impairment was the most important factor in classifying patients with schizophrenia versus healthy controls [48]. Using electroencephalographic patient data, Tikka and colleagues used SVM to classify patients with schizophrenia from healthy controls, and to classify patients with schizophrenia with positive symptoms from patients with schizophrenia with negative symptoms [49]. Particular eye movements are associated with schizophrenia and other disorders. However, it can be difficult to

experimentally obtain eye movement features and associate the movement pattern with a disease. Mao and colleagues in an exploratory study classified eye movements with a random forest of decision trees [50].

Work by Kim and colleagues published in 2020 trained a convolutional neural network to analyze social media users’ posts in a particular social media channel associated with a particular mental illness including schizophrenia, and thus create a deep learning model with natural language processing that could identify social media posts as belonging to various mental disorders including schizophrenia [51]. Work by Adler and colleagues published in 2020 used an app which collected schizophrenia patients’ passive cellphone data which included location, acceleration, app use, screen activity, text messages, and as well prompted users every 2–3 days to self-report positive and negative symptoms. A fully connected neural network autoencoder and recurrent unit sequence-to-sequence model learned input time series data from the patients, and was able to predict certain behavioral changes which occurred before there was a clinical relapse [52].

Yang and colleagues reviewed concepts in artificial intelligence that could help with the drug discovery process, although there is little discussion of potential medications which could be used with schizophrenia-related disorders [53]. Zilocchi and colleagues showed an interesting association of 245 mitochondrial proteins to bipolar disorder, schizophrenia, and mood disorders. The authors looked at mitochondrial proteins listed in a number of public databases and then deduced the proteins associated with bipolar disorder, schizophrenia, and mood disorders by comparisons with gene disease listings in the DisGeNET database which already compiles data from many repositories. The authors then examined pharmaceutical repositories for mood stabilizers and obtained a small number of active ingredients, which they then examined in a drug-gene interactions database, and found that seven of the active ingredients actually targeted a number of these 245 mitochondrial proteins. The authors suggested that future use of machine learning methods could allow a virtual synthesis process

that explored more of the chemical reactivity space and gave more possible potential drugs [54].

Work by Schneider took biological concepts from schizophrenia and applied them as constraints *back* to the field of artificial intelligence [55]. While neural networks, e.g., deep learning, have emerged as a very important technology in the field of artificial intelligence over the last decade, and while they can recognize patterns beyond human abilities, compared to a young child they are poor in logically and causally making sense of a problem they are solving. Deep learning networks have been getting better at performing their tasks by using ever faster computing hardware and training on ever larger sets of data, in a fashion that is not sustainable. As well, their amazing pattern recognition abilities do not give them many abilities we take for granted from human intelligence, e.g., the ability to usually explain why we made this or that decision.

Mammalian brains are characterized by cortex organized in repeating minicolumns. Schneider's Causal Cognitive Architecture 1 (CCA1) [56] pre-processes input sensory vectors through a hierarchy of Hopfield-like neural networks, but then feeds them to a navigation module. Unlike a deep neural network or unlike a symbolic artificial intelligence system, everything is processed and stored as navigation maps, which are inspired by the biological cortical minicolumns. Maps are stored, and retrieval can be triggered by other maps as well as operations on the maps. The architecture is not a tabula rasa but starts off with basic procedures termed "Instinctive Primitives" and learns procedures during its lifetime termed "Learned Primitives." The primitives are triggered by input sensory vectors and by maps in the navigation module. As well there are a number of other modules such as a "Goal/Emotion Module." This architecture is capable of performing pattern recognition much like artificial neural networks, but emerging almost automatically from the CCA1 architecture is a map-based precausal behavior to simple problems.

Just as feedback pathways abound in the human brain, they similarly exist in the CCA1 and downstream circuits can affect what inputs upstream circuits are to expect. The topic of

whether any animal, other than humans and possibly some primates, possesses true causal behavior is controversial, but no nonhuman possesses robust causal behavior, chimpanzees included. Of interest, psychosis seems to readily emerge in humans, e.g., even though only small percentage of the population will suffer from schizophrenia, more than 10% of the population will actually experience psychotic-like symptoms [57]. In Schneider's CCA1 architecture if the navigation module feeds back intermediate results to the input circuits, then they can be processed again, and over and over again, with each of the input-processing output cycles the architecture follows. Full causality readily emerges from the architecture when this occurs, as opposed to precausal behavior otherwise [56]. As well, the navigation maps do not need to be used for physical navigation but automatically allow navigation of complex concepts. As well, explainability emerges (which really is just sequential retrieval of maps executed), and the ability to use and create analogies almost automatically emerges from the architecture. However, in this fuller architecture, any small flaw of many possible such small flaws causes a cognitive dysfunction as well as misinterpretation of intermediate results as sensory input signals (i.e., delusion-like and hallucination-like).

Rezaii and colleagues used neural networks to show that during the prodromal phase, low semantic density, i.e., essentially what is clinically referred to as poverty of content, increased the likelihood of conversion to psychosis [58]. McFarlane and Illes discuss the work of Rezaii and other researchers who have developed techniques of identifying psychosis by machine learning methods of analyzing speech and use of social media, and write of the ethical concerns in making such early predictions of psychosis, some of which will be flawed [59].

Craig and colleagues, including Julian Leff, the inventor of the therapy, discuss patients with persecutory auditory hallucinations interacting with a digital avatar representing the hallucinated persecutor, such that the avatar becomes less hostile and the patient feels more in control. In a 12-week study of 75 patients receiving AVATAR therapy

versus controls receiving supportive therapy, there was a significant reduction in the severity of the auditory hallucinations in the AVATAR therapy group compared to the control group [60].

Work by Oh and colleagues, published in 2020, took ordinary structural MRI (1.5 and 3 T) brain images and trained a deep learning convolutional neural network to infer whether or not an image had cortical features of a patient with schizophrenia compared to normal subjects [61]. In the training data the neural network was able to correctly classify 840 out of 866 images, i.e., an accuracy of 97%. When the network was then tested on new images from different patients, the accuracy dropped, as is to be expected when a neural network starts classifying images it has not seen before in its training data, but it still was able to classify images reasonably well. It should be noted that human clinical specialists (five psychiatrists and two radiologists) had difficulty in discerning schizophrenia patients from the normal subjects in a random selection of the structural MRI images above.

Current and Future Clinical Use of AI Techniques in the Diagnosis and Treatment of Schizophrenia and Related Disorders

Fernandes and colleagues in 2017 write about what they consider the “new field of precision psychiatry” [62]. Just as “precision medicine” takes the individual features, i.e., the differences of each patient into account in crafting a prevention and treatment strategy [63], so does precision psychiatry. Bzdok and Meyer-Lindenberg write about the need for machine learning in precision psychiatry [64]. However, at the time of this writing, December 2020, the direct utilization of artificial intelligence techniques is not considered a standard of care (i.e., a legal term describing the expectations that health-care providers are to deliver in the care of their patients) in the diagnosis and treatment of patients with schizophrenia and schizophrenia-related disorders, in Canada where this chapter is being written, and it would seem worldwide as well. As well, in research

studies related to schizophrenia it could be argued that many studies incorporating artificial intelligence techniques could be rewritten to instead incorporate advanced conventional statistical techniques, including some of the references described above. However, all this is starting to change.

For example, consider the work by Koutsouleris and colleagues, with preliminary publication online in December 2020 [65]. In young people who meet the clinical high-risk (CHR) criteria for psychosis development, only about a fifth will have a transition to psychosis over a three-year timespan. As well, psychotic disorders do develop in individuals who would not have been classified in the CHR group. Koutsouleris and colleagues fed patient information including MRI imaging, clinical data, and neurocognitive data into a multimodal machine learning model, to predict the transition to a psychotic disorder in 167 patients meeting the criteria for CHR, 167 patients with a recent-onset depression, and 334 healthy matched controls. The mean age was 25.1 years old. The machine learning model was combined with clinicians’ risk estimates (where clinicians’ predictions had relatively high specificity but lower sensitivity versus the model’s relatively high sensitivity but lower specificity) and was able to predict the transition to psychosis with an accuracy of 85.9% (sensitivity 84.6%, specificity 87.3%). As a result, the authors recommend that this work be clinically implemented: “.... augmentation of human prognostic abilities with algorithmic pattern recognition improves prognostic accuracy to margins that likely justify the clinical implementation of cybernetic decision-support tools.”

At the time of this writing, there is much controversy about the need for functional neuroimaging in clinical as opposed to research psychiatry. However, in 2020 Henderson and colleagues [66] write about the field clinically using functional neuroimaging to “guide the ordering diagnostician to a better and more efficient evaluation and treatment of the neurobiological processes that underlie a particular patient’s symptoms.” As noted in many of the references above with regard to neuroimaging of

patients with possible schizophrenia, incorporation of AI techniques was a part of obtaining higher accuracies in the functional, as well as structural, imaging. For example, as noted above, Kalmady and colleagues [44] using AI techniques were able to differentiate healthy control subjects from patients with schizophrenia on fMRI with an accuracy of 87%. For example, as noted above, Oh and colleagues [61] using AI techniques were able to distinguish cortical features on ordinary structural MRIs between patients with schizophrenia and healthy controls. Indeed, Topol writes about the convergence of human decision-making with artificial intelligence in treating patients, particularly in the interpretation of imaging [67].

Starke and colleagues, writing in 2020, note that while machine learning is not routine in the clinical practice of psychiatry, in particular in schizophrenia, this is starting to occur, and thus there is a need to consider the ethical issues [68]. For example, machine learning algorithms, which could potentially be used in the future to diagnose and treat patients with schizophrenia, often do not explain well how they came up with their decisions.

The use of artificial intelligence in the clinical diagnosis and care of patients with schizophrenia is at the time of this writing in its infancy but starting to accelerate quickly. It is expected that in the 2020s not only will the capabilities of AI in the care of patients with schizophrenia keep improving, but the details of clinical implementations will start to become better managed.

Cross-References

- Applying Principles from Medicine Back to Artificial Intelligence

References

1. Russell S, Norvig P. What is AI? In: Artificial intelligence: a modern approach. 3rd ed. Upper Saddle River: Prentice Hall; 2010. p. 1–5.
2. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 5th ed. Arlington: American Psychiatric Association; 2013.
3. World Health Organization. The international statistical classification of diseases and related health problems, 10th revision (ICD-10). Geneva: World Health Organization; 2019. <https://icd.who.int/browse10/2019/en>
4. Hoffman RE. Computer simulations of neural information processing and the schizophrenia-mania dichotomy. *Arch Gen Psychiatry*. 1987;44:178–88.
5. Maurer K, Biehl K, Kühner C, Löffler W. On the way to expert systems. Comparing DSM-III computer diagnoses with CATEGO (ICD) diagnoses in depressive and schizophrenic patients. *Eur Arch Psychiatry Neurol Sci*. 1989;239(2):127–32. <https://doi.org/10.1007/BF01759586>.
6. Cohen JD, Servan-Schreiber D. Context, cortex, and dopamine: a connectionist approach to behavior and biology in schizophrenia. *Psychol Rev*. 1992;99(1): 45–77. <https://doi.org/10.1037/0033-295x.99.1.45>.
7. Hoffman RE, McGlashan TH. Parallel distributed processing and the emergence of schizophrenic symptoms. *Schizophr Bull*. 1993;19(1):119–40. <https://doi.org/10.1093/schbul/19.1.119>.
8. Garfield DA, Rapp C. Application of artificial intelligence principles to the analysis of “crazy” speech. *J Nerv Ment Dis*. 1994;182(4):205–11. <https://doi.org/10.1097/00005053-199404000-00002>.
9. Seeman MV. Neural networks and schizophrenia. *Can J Psychiatr*. 1994;39(8):353–4. <https://doi.org/10.1177/070674379403900801>.
10. Lowell WE, Davis GE. Predicting length of stay for psychiatric diagnosis-related groups using neural networks. *J Am Med Inform Assoc*. 1994;1(6):459–66. <https://doi.org/10.1136/jamia.1994.95153435>.
11. Hoffman RE, McGlashan TH. Reduced corticocortical connectivity can induce speech perception pathology and hallucinated ‘voices’. *Schizophr Res*. 1998;30(2): 137–41. [https://doi.org/10.1016/s0920-9964\(97\)00142-4](https://doi.org/10.1016/s0920-9964(97)00142-4).
12. Corson PW, Nopoulos P, Andreasen NC, Heckel D, Arndt S. Caudate size in first-episode neuroleptic-naïve schizophrenic patients measured using an artificial neural network. *Biol Psychiatry*. 1999;46(5):712–20. [https://doi.org/10.1016/s0006-3223\(99\)00079-7](https://doi.org/10.1016/s0006-3223(99)00079-7).
13. Razzouk D, Mari JJ, Shirakawa I, Wainer J, Sigulem D. Decision support system for the diagnosis of schizophrenia disorders. *Braz J Med Biol Res*. 2006;39(1): 119–28. <https://doi.org/10.1590/s0100-879x20060000100014>.
14. Jafri MJ, Calhoun VD. Functional classification of schizophrenia using feed forward neural networks. *Conf Proc IEEE Eng Med Biol Soc*. 2006; Suppl:6631–4. <https://doi.org/10.1109/IEMBS.2006.260906>.
15. Bose SK, Turkheimer FE, Howes OD, Mehta MA, Cunliffe R, Stokes PR, Grasby PM. Classification of schizophrenic patients and healthy controls using [¹⁸F]

- fluorodopa PET imaging. *Schizophr Res.* 2008;106(2–3):148–55. <https://doi.org/10.1016/j.schres.2008.09.011>.
16. Struyf J, Dobrin S, Page D. Combining gene expression, demographic and clinical data in modeling disease: a case study of bipolar disorder and schizophrenia. *BMC Genomics.* 2008;9:531. <https://doi.org/10.1186/1471-2164-9-531>.
17. Ozyurt IB, Brown GG. Knowledge discovery via machine learning for neurodegenerative disease researchers. *Methods Mol Biol.* 2009;569:173–96. https://doi.org/10.1007/978-1-59745-524-4_9.
18. Eliasmith C, Anderson CH. Neural engineering: representation, computation, and dynamics in neurobiological systems. Cambridge, MA: MIT Press; 2003.
19. Siekmeier PJ. Evidence of multistability in a realistic computer simulation of hippocampus subfield CA1. *Behav Brain Res.* 2009;200(1):220–31. <https://doi.org/10.1016/j.bbr.2009.01.021>.
20. Karolidis DA, Agatonovic-Kustrin S, Morton DW. Artificial neural network (ANN) based modelling for D1 like and D2 like dopamine receptor affinity and selectivity. *Med Chem.* 2010;6(5):259–70. <https://doi.org/10.2174/157340610793358891>.
21. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th international conference on neural information processing systems – volume 1 (NIPS'12). Red Hook: Curran Associates; 2012.
22. Veronese E, Castellani U, Peruzzo D, Bellani M, Brambilla P. Machine learning approaches: from theory to application in schizophrenia. *Comput Math Methods Med.* 2013;2013:867924. <https://doi.org/10.1155/2013/867924>.
23. Kim J, Calhoun VD, Shim E, Lee JH. Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. *Neuroimage.* 2016;124(Pt A):127–46. <https://doi.org/10.1016/j.neuroimage.2015.05.018>.
24. Arbabshirani MR, Plis S, Sui J, Calhoun VD. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. *Neuroimage.* 2017;145(Pt B):137–65. <https://doi.org/10.1016/j.neuroimage.2016.02.079>.
25. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci Rep.* 2016;6:26094. <https://doi.org/10.1038/srep26094>.
26. Raffard S, Bortolon C, Khoramshahi M, Salesse RN, Burca M, Marin L, Bardy BG, Billard A, Macioce V, Capdevielle D. Humanoid robots versus humans: how is emotional valence of facial expressions recognized by individuals with schizophrenia? An exploratory study. *Schizophr Res.* 2016;176(2–3):506–13. <https://doi.org/10.1016/j.schres.2016.06.001>.
27. Arnon S, Dahan N, Koren A, Radiano O, Ronen M, Yannay T, Giron J, Ben-Ami L, Amir Y, Hel-Or Y, Friedman D, Bachelet I. Thought-controlled nanoscale robots in a living host. *PLoS One.* 2016;11(8):e0161227. <https://doi.org/10.1371/journal.pone.0161227>.
28. Mikolas P, Melicher T, Skoch A, Matejka M, Slováková A, Bakstein E, Hajek T, Spaniel F. Connectivity of the anterior insula differentiates participants with first-episode schizophrenia spectrum disorders from controls: a machine-learning study. *Psychol Med.* 2016;46(13):2695–704. <https://doi.org/10.1017/S0033291716000878>.
29. Zarogianni E, Storkley AJ, Johnstone EC, Owens DG, Lawrie SM. Improved individualized prediction of schizophrenia in subjects at familial high risk, based on neuroanatomical data, schizotypal and neurocognitive features. *Schizophr Res.* 2017;181:6–12. <https://doi.org/10.1016/j.schres.2016.08.027>.
30. Pinaya WH, Gadelha A, Doyle OM, Noto C, Zugman A, Cordeiro Q, Jackowski AP, Bressan RA, Sato JR. Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia. *Sci Rep.* 2016;6:38897. <https://doi.org/10.1038/srep38897>.
31. Pergola G, Trizio S, Di Carlo P, Taurisano P, Mancini M, Amoroso N, Nettis MA, Andriola I, Caforio G, Popolizio T, Rampino A, Di Giorgio A, Bertolino A, Blasi G. Grey matter volume patterns in thalamic nuclei are associated with familial risk for schizophrenia. *Schizophr Res.* 2017;180:13–20. <https://doi.org/10.1016/j.schres.2016.07.005>.
32. Keshavan MS, Sudarshan M. Deep dreaming, aberrant salience and psychosis: connecting the dots by artificial neural networks. *Schizophr Res.* 2017;188:178–81. <https://doi.org/10.1016/j.schres.2017.01.020>.
33. Bain EE, Shafner L, Walling DP, Othman AA, Chuang-Stein C, Hinkle J, Hanina A. Use of a novel artificial intelligence platform on mobile devices to assess dosing compliance in a phase 2 clinical trial in subjects with schizophrenia. *JMIR Mhealth Uhealth.* 2017;5(2):e18. <https://doi.org/10.2196/mhealth.7030>.
34. Birnbaum ML, Ernala SK, Rizvi AF, De Choudhury M, Kane JM. A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. *J Med Internet Res.* 2017;19(8):e289. <https://doi.org/10.2196/jmir.7956>.
35. Dluhoš P, Schwarz D, Cahn W, van Haren N, Kahn R, Španiel F, Horáček J, Kašpárek T, Schnack H. Multi-center machine learning in imaging psychiatry: a meta-model approach. *NeuroImage.* 2017;155:10–24. <https://doi.org/10.1016/j.neuroimage.2017.03.027>.
36. Honnorat N, Dong A, Meisenzahl-Lechner E, Koutsouleris N, Davatzikos C. Neuroanatomical heterogeneity of schizophrenia revealed by semi-supervised machine learning methods. *Schizophr Res.* 2019;214:43–50. <https://doi.org/10.1016/j.schres.2017.12.008>.
37. Winterburn JL, Voineskos AN, Devenyi GA, Plitman E, de la Fuente-Sandoval C, Bhagwat N, Graff-Guerrero A, Knight J, Chakravarty MM. Can we accurately classify schizophrenia patients from

- healthy controls using magnetic resonance imaging and machine learning? A multi-method and multi-dataset study. *Schizophr Res.* 2019;214:3–10. <https://doi.org/10.1016/j.schres.2017.11.038>.
38. Zeng LL, Wang H, Hu P, Yang B, Pu W, Shen H, Chen X, Liu Z, Yin H, Tan Q, Wang K, Hu D. Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity MRI. *EBioMedicine*. 2018;30:74–85. <https://doi.org/10.1016/j.ebiom.2018.03.017>.
 39. Bae Y, Kumarasamy K, Ali IM, Korfiatis P, Akkus Z, Erickson BJ. Differences between schizophrenic and normal subjects using network properties from fMRI. *J Digit Imaging*. 2018;31(2):252–61. <https://doi.org/10.1007/s10278-017-0020-4>.
 40. Mikolas P, Hlinka J, Skoch A, Pitra Z, Frodl T, Spaniel F, Hajek T. Machine learning classification of first-episode schizophrenia spectrum disorders and controls using whole brain white matter fractional anisotropy. *BMC Psychiatry*. 2018;18(1):97. <https://doi.org/10.1186/s12888-018-1678-y>.
 41. Zhao K, So HC. Drug repositioning for schizophrenia and depression/anxiety disorders: a machine learning approach leveraging expression data. *IEEE J Biomed Health Inform.* 2019;23(3):1304–15. <https://doi.org/10.1109/JBHI.2018.2856535>.
 42. Lin E, Lin CH, Lai YL, Huang CH, Huang YJ, Lane HY. Combination of G72 genetic variation and G72 protein level to detect schizophrenia: machine learning approaches. *Front Psych*. 2018;9:566. <https://doi.org/10.3389/fpsyg.2018.00566>.
 43. Fond G, Bulzacka E, Boucckine M, Schürhoff F, Berna F, Godin O, Aouizerate B, Capdevielle D, Chereau I, D'Amato T, Dubertret C, Dubreucq J, Faget C, Leignier S, Lançon C, Mallet J, Misrahi D, Passerieu C, Rey R, Schandrin A, Urbach M, Vidailhet P, Leboyer M, FACE-SZ (FondaMental Academic Centers of Expertise for Schizophrenia) Group, Boyer L, Llorca PM. Machine learning for predicting psychotic relapse at 2 years in schizophrenia in the national FACE-SZ cohort. *Prog Neuropsychopharmacol Biol Psychiatry*. 2019;92:8–18. <https://doi.org/10.1016/j.pnpbp.2018.12.005>.
 44. Kalmary SV, Greiner R, Agrawal R, Shivakumar V, Narayanaswamy JC, Brown MRG, Greenshaw AJ, Dursun SM, Venkatasubramanian G. Towards artificial intelligence in mental health by improving schizophrenia prediction with multiple brain parcellation ensemble-learning. *NPJ Schizophr*. 2019;5(1):2. <https://doi.org/10.1038/s41537-018-0070-8>.
 45. Brodsky BB, Girgis RR, Favorov OV, Bearden CE, Woods SW, Addington J, Perkins DO, Walker EF, Cornblatt BA, Brucato G, Purcell SE, Brodsky IS, Cadenehead KS. The Early Psychosis Screener for Internet (EPSI)-SR: predicting 12 month psychotic conversion using machine learning. *Schizophr Res.* 2019;208: 390–6. <https://doi.org/10.1016/j.schres.2019.01.015>.
 46. Barrera A, Gee C, Wood A, Gibson O, Bayley D, Geddes J. Introducing artificial intelligence in acute psychiatric inpatient care: qualitative study of its use to conduct nursing observations. *Evid Based Ment Health*. 2020;23(1):34–8. <https://doi.org/10.1136/ebmental-2019-300136>.
 47. Wu CS, Luedtke AR, Sadikova E, Tsai HJ, Liao SC, Liu CC, Gau SS, VanderWeele TJ, Kessler RC. Development and validation of a machine learning individualized treatment rule in first-episode schizophrenia. *JAMA Netw Open*. 2020;3(2):e1921660. <https://doi.org/10.1001/jamanetworkopen.2019.21660>.
 48. Parola A, Salvini R, Gabbatore I, Colle L, Berardinelli L, Bosco FM. Pragmatics, theory of mind and executive functions in schizophrenia: disentangling the puzzle using machine learning. *PLoS One*. 2020;15(3):e0229603. <https://doi.org/10.1371/journal.pone.0229603>.
 49. Tikka SK, Singh BK, Nizamie SH, Garg S, Mandal S, Thakur K, Singh LK. Artificial intelligence-based classification of schizophrenia: a high density electroencephalographic and support vector machine study. *Indian J Psychiatry*. 2020;62(3):273–82. https://doi.org/10.4103/psychiatry.IndianJPsycho_91_20.
 50. Mao Y, He Y, Liu L, Chen X. Disease classification based on eye movement features with decision tree and random forest. *Front Neurosci*. 2020;14:798. <https://doi.org/10.3389/fnins.2020.00798>.
 51. Kim J, Lee J, Park E, Han J. A deep learning model for detecting mental illness from user content on social media. *Sci Rep*. 2020;10(1):11846. <https://doi.org/10.1038/s41598-020-68764-y>.
 52. Adler DA, Ben-Zeev D, Tseng VW, Kane JM, Brian R, Campbell AT, Hauser M, Scherer EA, Choudhury T. Predicting early warning signs of psychotic relapse from passive sensing data: an approach using encoder-decoder neural networks. *JMIR Mhealth Uhealth*. 2020;8(8):e19962. <https://doi.org/10.2196/19962>.
 53. Yang X, Wang Y, Byrne R, Schneider G, Yang S. Concepts of artificial intelligence for computer-assisted drug discovery. *Chem Rev*. 2019;119(18): 10520–94. <https://doi.org/10.1021/acs.chemrev.8b00728>.
 54. Zilocchi M, Broderick K, Phanse S, Aly KA, Babu M. Mitochondria under the spotlight: on the implications of mitochondrial dysfunction and its connectivity to neuropsychiatric disorders. *Comput Struct Biotechnol J*. 2020;18:2535–46. <https://doi.org/10.1016/j.csbj.2020.09.008>.
 55. Schneider H. The meaningful-based cognitive architecture model of schizophrenia. *Cogn Syst Res*. 2020;59:73–90. <https://doi.org/10.1016/j.cogsys.2019.09.019>.
 56. Schneider H. Causal cognitive architecture 1: integration of connectionist elements into a navigation-based framework. *Cogn Syst Res*. 2021;66:67–81. <https://doi.org/10.1016/j.cogsys.2020.10.021>.
 57. van Os J, Hanssen M, Bijl RV, et al. Prevalence of psychotic disorder and community level psychotic symptoms: an urban-rural comparison. *Arch Gen Psychiatry*. 2001;58(7):663–8.

58. Rezaii N, Walker E, Wolff P. A machine learning approach to predicting psychosis using semantic density and latent content analysis. *npj Schizophr.* 2019;5:9. <https://doi.org/10.1038/s41537-019-0077-9>.
59. McFarlane J, Illes J. Neuroethics at the interface of machine learning and schizophrenia. *npj Schizophr.* 2020;6:18. <https://doi.org/10.1038/s41537-020-0108-6>.
60. Craig TKJ, Rus-Calafell M, Ward T, Leff JP, Huckvale M, Howarth E, Emsley R, Garety PA. AVATAR therapy for auditory verbal hallucinations in people with psychosis: a single-blind, randomised controlled trial. *Lancet Psychiatry.* 2018;5(1):31–40. [https://doi.org/10.1016/S2215-0366\(17\)30427-3](https://doi.org/10.1016/S2215-0366(17)30427-3).
61. Oh J, Oh B-L, Lee K-U, Chae J-H, Yun K. Identifying schizophrenia using structural MRI with a deep learning algorithm. *Front Psych.* 2020;11:16. <https://doi.org/10.3389/fpsyg.2020.00016>.
62. Fernandes BS, Williams LM, Steiner J, Leboyer M, Carvalho AF, Berk M. The new field of ‘precision psychiatry’. *BMC Med.* 2017;15(1):80. <https://doi.org/10.1186/s12916-017-0849-x>.
63. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med.* 2015;372(9):793–5. <https://doi.org/10.1056/NEJMmp1500523>.
64. Bzdok D, Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. *Biol Psychiatry Cogn Neurosci Neuroimaging.* 2018;3(3):223–30. <https://doi.org/10.1016/j.bpsc.2017.11.007>.
65. Koutsouleris N, Dwyer DB, Degenhardt F, et al. Multimodal machine learning workflows for prediction of psychosis in patients with clinical high-risk syndromes and recent-onset depression. *JAMA Psychiatry.* Published online December 02, 2020. <https://doi.org/10.1001/jamapsychiatry.2020.3604>.
66. Henderson TA, van Lierop MJ, McLean M, et al. Functional neuroimaging in psychiatry—aiding in diagnosis and guiding treatment. What the American Psychiatric Association does not know. *Front Psych.* 2020;11:276. <https://doi.org/10.3389/fpsyg.2020.00276>.
67. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25:44–56. <https://doi.org/10.1038/s41591-018-0300-7>.
68. Starke G, De Clercq E, Borgwardt S, Elger BS. Computing schizophrenia: ethical challenges for machine learning in psychiatry. *Psychol Med.* 2020;15:1–7. <https://doi.org/10.1017/S0033291720001683>.



The Rise of the Mental Health Chatbot

115

Michiel Rauws

Contents

Introduction	1610
Mental Health Support using Chatbots	1610
Chapter Summary	1610
Mental Health Chatbots	1610
The Value of Mental Health Chatbots	1610
How Chatbots Work	1611
Economic Impact	1611
How a Chatbot Could Change the Economics of the Employee Benefits Industry ...	1611
Real-World Applications	1613
Case Study: 24/7 Access for 430,000 People Served by a Public Health Department	1613
Case Study: AI Mental Health Support for Young Mothers in Africa	1614
Case Study: Supporting Caregivers and Their Patients	1616
Case Study: Addressing Comorbidities in Childhood Obesity, Prediabetes, and Mental Health Struggles	1617
References	1618

Abstract

The most evident use of artificial intelligence within the mental health field are within chatbots. The adoption of mental health chatbots is a nascent field with the potential to fill mental health treatment gaps and alter perceptions of support to help reduce depression and anxiety in patients. This chapter will define mental health chatbots and explore

their measurable benefits through applications in case studies and an analysis of technological opportunities within the health care industry.

Keywords

Chatbot · Mental health · Coaching · Wellness · Health care · Behavioral health · Benefits · Health insurance · Self-insured · Depression · Anxiety

M. Rauws (✉)
X2 AI, San Francisco, CA, USA
e-mail: michiel@x2ai.com

Introduction

The most evident use of artificial intelligence within the mental health field are within chatbots. The adoption of mental health chatbots is a nascent field with the potential to fill mental health treatment gaps and alter perceptions of support to help reduce depression and anxiety in patients. This chapter will define mental health chatbots and explore their measurable benefits through applications in case studies with an analysis of the* technological opportunities within the health care industry.

Mental Health Support using Chatbots

Chapter Summary

The most evident use of artificial intelligence within the mental health field are within chatbots. The adoption of mental health chatbots is a nascent field with the potential to fill mental health treatment gaps and alter perceptions of support to help reduce depression and anxiety in patients. This chapter will define mental health chatbots and explore their measurable benefits through applications in case studies and an analysis of technological opportunities within the health care industry.

Mental Health Chatbots

The first AI chatbot, ELIZA, was developed at MIT and used natural language processing to conduct mock psychotherapy chats with people. Due to the limited availability of personal computers and cell phones, ELIZA was never turned into a commercial product. Where ELIZA was limited, the capabilities of modern chatbots have expanded rapidly thanks to technological advancements in AI memory and emotion identification.

Modern-day artificial intelligence chatbots are computer programs designed to simulate how a human would behave as a conversational partner and encourages* authentic exchanges with patients. Chatbots are a low-cost, user-friendly, and highly customizable solution that enable emotional support to be scaled to hundreds of

thousands of people at a time. Accessing support from a chatbot is convenient through existing and familiar communication channels, including text messaging (iMessage or SMS), Facebook Messenger, and integrations with Amazon Alexa/Google Home voice-enabled services [1].

The Value of Mental Health Chatbots

Mental illness impacts many facets of life. Within the workforce, 6.8% of people are directly impacted by depression and as a result \$11,936 is lost annually on average per affected employee due to absenteeism, disability, and lack of productivity. For many people, access to mental health support is hampered by long wait times, high treatment costs, as well as the social stigma in seeking mental health support. Nearly 60% of people with a mental health problem encounter barriers to receiving care.

Chatbots can lower these barriers to access by providing flexible and scalable support to the user. Designed to have conversations with thousands of people a day, chatbots can support underserved groups through on-demand 24/7 availability and eliminate travel costs or the need to schedule appointments. Individuals may find it challenging to talk about mental health, and may be more likely to share and disclose information using chatbots than through traditional in-person therapy, enhancing a sense of privacy and quelling stigmatic concerns.

In 2011, it was estimated that 37% of individuals on Medicare had a serious mental health disorder [2]. Mood disorders are associated with relatively high treatment center readmission rates, further taxing the already limited availability within the inpatient mental health system. Within 30 days, 15% of patients with mood disorders are readmitted nationally and across the USA; the average cost for the repeated hospital stay is approximately \$7200 [3].

A key factor in influencing the rate of readmission includes poor outpatient follow-up [4, 5]. Follow-up calls often follow a structured script which a call center nurse follows, but through a chatbot these call scripts can be converted to chatbot scripts. Through the

electronic health record (EHR), a digital version of a patient's paper chart, patients can opt in to allow a chatbot follow-up by sending them a text message. When a predetermined script is not sufficient for the needs of the conversation, it is escalated to call center staff to take over the conversation from the chatbot. The majority of these conversations can be handled by the chatbot which provides increased access to care and a reduction in the* total cost of care.

How Chatbots Work

To explain how chatbots work we focus on one case study analyzing Tess, a chatbot developed by X2AI Inc. used by health care systems, employers, and employee assistance programs (EAP's) to deliver emotional wellness support. Tess is designed using a combination of technologies, emotional algorithms, and artificial intelligence techniques to learn about the needs of the user and offer personalized support. In collaboration with mental health professionals, Tess repeats conversations that specialists have crafted, in order to engage with users in an empathetic way that is appropriate to the inputted emotion or scenario. All recommendations from Tess require sign-off from a counselor to ensure safety and consistent quality support.

Specific interventions are employed based on the patient's reported concern. If the patient tells Tess that they feel anxious, Tess may offer a strategy to help them achieve a more relaxed state. When starting a conversation, Tess states "It's important for you to know that if you are in a crisis you should contact emergency services. You can type SOS for resources after our first chat is over. OK?" Anytime a user types or says "SOS" or reports suicidal ideation during the conversation, Tess offers resources. Tess also provides basic risk assessment and sends crisis alerts to a counselor or crisis center, instructing humans to take over the conversation in order to de-escalate the situation. Similarly, to how a caseworker would make recommendations on patient care, chatbots can also leverage existing resources and refer patients to other digital health tools.

There are different ways to build mental health chatbots. Companies like X2AI can provide a customized version of an existing chatbot, and platform that allows clients to build chatbots from the ground up. There are general purpose tools available in the market as well that can build chatbots from the ground up, such as Google Dialogflow or Amazon Lex. Building a bot from the ground up requires* significant amounts of resources and time, therefore most companies elect to customize an existing chatbot.

The Tess system has a conversation library of 800 different supportive interventions, consisting of* over 3,000,000 unique conversations. Over 50 coaches and psychologists have engineered the conversations available through Tess and these conversations are personalized to each user's needs to deliver a wide range of emotional wellness intervention—from cognitive behavioral therapy to deeper, psychodynamic strategies. It is vital that the authoring and crafting of these conversations are done according to high-quality standards. These conversations are used by thousands of people who chat with the chatbot. Ethical considerations need to be part of the quality control process, for example, in order to ensure the content will not include any bias [6].

Tess processes quantitative and qualitative feedback from users to automatically enhance the system and improve the user experience. Conversations with the highest helpfulness ratings—measured by Tess asking the user about the helpfulness of the conversation after every intervention—are prioritized to be delivered before others to ensure relevancy and utility. Through each conversation Tess tailors the frequency of conversation initiation, timing of check-ins, and increases the capacity to identify emotions.

Economic Impact

How a Chatbot Could Change the Economics of the Employee Benefits Industry

Employers purchase an employee assistance program (EAP) for their employees as part of a benefits package. These programs offer extensive

benefits ranging from mental health support, personal legal assistance to financial planning. The potential for change in the economics of this niche industry can illustrate how the health care industry can toward technology adoption.

Historically, a high level of trust existed between EAP providers and employers. Employees were pleased with the introduction of new support offerings, and utilization was measured by the hours clinicians billed and considered successful at 15–25%. As a result, employers happily shared the benefits of better rates as EAP providers discovered lower rates as they bundle demand; a \$10 price per employee per month (PEPM) was sustainable and led to cost savings for employers overall.

A downward trend in both utilization levels and PEPM have created a vicious cycle, due to numerous reasons we do not have bandwidth to cover in this chapter. Currently the industry-wide average utilization hovers around 5%, with a PMPM as low as \$1.25. Several factors which maintain this balance are starting to change. First, we will list the factors that have contributed to the status quo. Then, we will explore which of these strategies can be used to counteract the negative effect of these factors. Finally, we discuss practical takeaways and considerations on how to implement these strategies in order for EAP's and eventually the industry as a whole—to shift back to higher utilization and higher PEPM rates.

Factors Causing Low Utilization

1. A low PEPM leaves less marketing budget. Limited resources allocated to marketing makes it hard to reach and engage the employee population. Benefit fairs require travel and the* hourly costs of staff being at the fair.
2. The EAP offering does not often match employee needs. If an EAP only offers face-to-face sessions, then individuals who could be helped with low-cost self-help interventions will instead use costly face-to-face sessions. Higher costs of service delivery results* in less budget for marketing to drive utilization.
3. Face-to-face counseling is costly, and thus there has to be a strict eligibility check

procedures* in place. Accessing online self-help resources require an eligibility check in order to track utilization. Requirements such as entering a username and password to access any EAP resource serve as barriers for those in need.

Utilization Strategies of Change

1. Partner with an innovative startup. The way the marketing dollars are deployed can make a big difference in the utilization that is achieved. Therefore marketing budgets can be expected to remain the same. Informing members about their benefits through social media creates a direct relationship between the member and the EAP, making these benefits more accessible. The standard way of improving social media marketing returns is to pay for a professional marketing firm, or provide time-consuming social media training for current staff members. To test if such an investment is worthwhile for your population, the most cost-effective way could be to work with an innovative technology provider. These companies offer digital products to members and have a low marginal cost, allowing their products to be used as much as possible.
2. Cut costs and scale support with a digital mental health solution. Face-to-face sessions are the most trusted way of providing support for mental health issues. For many EAPs this is also the only option offered to get support with any mental health issues. However, a large percentage of members might not need any face-to-face sessions in order to cope. Sometimes a conflict with a coworker can lead to high levels of anxiety, for which a digital intervention could suffice. Wait times for face-to-face sessions often range from a few days to many weeks, whereas digital interventions are available on demand. In this way, members get easier access to care, while EAPs save funds as digital interventions cost a fraction of face-to-face sessions.
3. Streamline the eligibility process. A strict eligibility policy brings down utilization and drives costs for the EAP call center, as members need support to retrieve their login

credentials, or verify while on the phone. However, services which have a lower marginal cost, such as blog posts and informational content, do not need to be accessed only after an eligibility check. Alternatively, tracking utilization can also be done through the use of a chatbot. The chatbot only needs to do a verification once and afterward it recognizes the member by phone number. Then, when the chatbot recommends a blog post or an article, the utilization gets tracked automatically.

Factors Causing Low PEPM

1. Measuring utilization is easy, which provides the purchasers with a very strong negotiation position to demand lower PEPM rates.
2. Measuring ROI is hard, and research is often not generalizable. Results from research studies of EAP programs on symptom reduction and ROI are often not generalizable as specific employer populations are very diverse. It is costly to measure the impact of services offered, and either has to be done by a clinician or a nurse calling from a call center to take necessary assessments. The financial burden of this process limits funding toward research studies for each specific employee population.
3. Purchasers spend their budget on procuring counseling services directly from providers. These providers charge a fee for service, and use substantial marketing budgets to engage the population. This takes away from the PEPM budget for EAPs.

PEPM Strategies of Change

1. Use a chatbot to measure engagement. Due to the eligibility check discussed above, it is hard to show high utilization on the offering that would have been used more. Using a chatbot allows you to track exactly how much time members spend interacting with your content.
2. Reduce ROI research costs with a digital solution. ROI studies are costly, due to the cost of the intervention itself, the cost of recruitment, and gathering data. Participant recruitment and the gathering of data is often done over the phone through the call center which is expensive, whereas e-mail is not a feasible way to get

enough engagement. Digital interventions are much more accessible and can deliver surveys and questionnaires at the members convenience such as before starting the next chapter in the online course, or through a check-in from a chatbot. Digital interventions make it possible to gather pre- and post-intervention measurements, and can intervene itself as well. This fully automated research process brings down the cost of an ROI study, which makes it possible to create a custom study for every account or employee population.

3. End the competition, and start collaborating. New players in the market provide their data on high levels of engagement which are powered by large marketing budgets, and the use of social media strategies as mentioned above. In this way they compete with existing EAPs for the budget of the purchaser. In the short term, these players are able to get market share. In the long term, growth will stagnate as they become subject to the same overhead as EAPs are facing. A more efficient approach for both sides is not to compete but to partner in providing the member better access to care.

Real-World Applications

To demonstrate the impact of chatbots in real-world applications, the pages below show how versatile chatbots are and describe a wide array of different use cases.

Case Study: 24/7 Access for 430,000 People Served by a Public Health Department

This Public Health Department manages a population of 430,000 people in a region with an agricultural output of over 4 billion dollars, and provides access to mental health care to anyone in need regardless of insurance coverage or citizenship status. Many seasonal workers who come to the region are essential to the local food production industry, but these jobs are low paying and generally do not come with insurance benefits.

The Latino immigrant population greatly depends on the health department to get access to care and the demand for Spanish-speaking psychologists is high. Due to budget constraints and a shortage of counselors in the region, wait times for support were often higher than 6 months to a year. Outsourcing part of the mental health treatments to traditional behavioral health providers misaligned incentives, increased costs, and was unable to resolve barriers to accessing care. Compounded with a high demand for care in the region, the workload for existing staff contributed to an overtaxed system and feelings of burnout.

Through state-funded grant studies, the public health department found that increasing free access to care effectively lowered the total cost of care for the region's population. Through a strategic population management approach in an effort to provide quality care at an affordable cost, Public Health Department made Tess available free of charge to anyone within the region in both English and Spanish, and users could refer friends and family.

Tess is a mental health text support service available through toll-free SMS and Facebook Messenger. Similar services rely on trained counselors to exchange messages with patients, but are limited by the amount of people you can reach due to budget, staff bandwidth, and hiring constraints. Tess offers the same text support chat model but instead is automated by an artificial intelligence system which repeats pre-written mental health scripts and learns from each conversation.

To alleviate initial concerns in implementing a chatbot, a lightweight integration was created to exchange data on usage metrics to case managers. Additionally, whenever a person was in crisis and a conversation needed to be escalated, human counselors were able to take over the conversation and notify the health department of the incident.

By implementing Tess as a resource in the community, the public health department was able to increase the quality of care by offering 24/7 on-demand mental health support. The total cost of care was affected in two ways: [1] directly lowering behavioral health care costs and [2] indirectly lowering total cost of care through

increased access to behavioral health support. Traditionally, this level of care would normally cost \$2200 per treatment, whereas the use of Tess only costs \$5 per person per month.

Tess facilitated a significant increase in access to care and clients experienced a 100% decrease in wait times due to the on-demand nature of the program. In general, when a population gets access to Tess the following impact is observed. Eighty-seven point five percent move toward recovery with a reduction in symptoms of up to 50%. Additionally, the remaining 12.5% of the population sees a decrease in the severity of symptoms by 50–100%. The benchmark for a successful treatment by a counselor is a reduction of 50% in the severity of symptoms or more, proving that—in some cases—the chatbot is as effective as its human counterparts.

Case Study: AI Mental Health Support for Young Mothers in Africa

Original article: <https://www.x2ai.com/blog/duke-university-postpartum-depression-support>

In 2019 Duke University launched a new version of their Healthy Moms program, bolstered with an AI-powered on-demand chatbot, to help women navigate pre- and postpartum mental health concerns in Kenya. Both as a new form of treatment and in combination with traditional treatments, the program focused on increasing the well-being of mothers in underserved populations by reducing the chances of developing pre- and postpartum depression. Inaccessibility to care can be caused by a multitude of factors, including socioeconomic status, race/ethnicity, and demographics.

Lack of access to care affects both women and their children, leading to increased maternal morbidity and mortality, poor infant health, and poor developmental outcomes. Most cases of depression in low-to-middle-income countries go untreated due to a lack of trained professionals to assist with pregnancies and the postpartum period. Outside of Nairobi, Kenya, a staggering ratio of 1 provider per 200,000–250,000 people provides professional care.

In an effort to close this gap in care, the World Health Organization developed the Mental Health Gap Action Program (mhGAP) intervention guide to assist nonspecialist providers address in providing mental health services in primary health care settings. The Thinking Healthy Program, an example of mhGAP, is a 15-session, cognitive behavior therapy (CBT)-based intervention for treating perinatal depression. This is typically done by community health workers with no specific background in mental health. They are trained to help pregnant women identify unhealthy thinking and equip them with the resources to practice thinking and acting healthy.

Despite the potential of such programs, most women in low-to-middle-income countries who need treatment still do not have access to care due to the common complications most task-sharing models face, such as lack of funding and infrastructure. By adapting the Thinking Healthy program to be delivered through AI technology like Zuri, a chatbot created by X2, Duke University aimed to break down those barriers in low-to-middle-income countries by making it possible for anyone with a basic phone to receive high-quality, psychological support at any time, regardless of location. Duke University's Healthy Moms program utilized Zuri to deliver the mhGAP Thinking Healthy Program via technological adaptation.

A single-case experimental design using Zuri was conducted with pregnant women and new mothers recruited from hospitals outside of Nairobi, Kenya. Eligible participants would have regular conversations modeled from the Thinking Healthy program via Zuri and asked women to check in every 3 days during the 1- to 2-week randomized baseline and interventions periods to evaluate progress.

Zuri would ask participants to report their mood, asking questions such as "How are you feeling now?" If the response was positive or neutral, Zuri would respond by offering a coaching chat such as music, cooking, or passions. If the response was a negative emotion, Zuri would respond with supportive intervention such as mindfulness or relaxation. There was no limit to the number of conversations or how often

a participant wanted to communicate with Zuri. All modules, both positive and negative, were prioritized based on combined helpfulness ratings in which women reported on the utility of the modules.

Participants expressed that they credited Zuri for positive life changes based on their conversations, and that they trusted this AI technology because it was more unbiased and provided both factual and useful information. Many women said that they preferred to chat with Zuri than with a counselor because they felt they could be more open with a chatbot. When the mothers were asked what their favorite part of Healthy Moms was, many attested to the easy and relaxing exercises learned through Zuri, including meditation, breathing, and walking. They also raved about the helpful advice they received from the Healthy Mom program such as breastfeeding and how to play with the child.

Emotional well-being AI programs have great potential to address the large treatment gaps that exist in underserved populations and improve moods when used both as a new form of treatment and in combination with traditional treatments. Zuri works by engaging a patient in conversation through text messaging. Either Zuri or the patient can start a conversation, and Zuri walks the patient through a structured curriculum, like the one defined in the Think Healthy program. Zuri also has a safety measure built in to engage human counselors if additional support is needed. AI-enhanced systems such as Zuri improve over time with practice and experience just like mental health specialists and nonspecialists trained to deliver psychotherapy. Over time, Zuri's emotional recognition algorithm was updated when it interpreted the emotional responses of the participant. Adding AI chatbots as a form of care allows convenient and cost-effective interactions to keep patients engaged and on track throughout their pregnancy. It also adds an extra layer of care and support outside of office hours to ensure the mother-to-be has the care she needs to maintain her mental health.

It was concluded that Zuri was a valuable resource accepted by the sample pregnant women and new mothers participating in the

beta launch of Healthy Moms. The participant's mood improved by 7% over time. The next steps to making Zuri a more acceptable and usable platform will be to explore different channels to deliver and communicate, refine intervention conversation flow, and build Zuri into a multilingual platform to include Swahili to better support the user demographic [7].

Case Study: Supporting Caregivers and Their Patients

Chatbot technology is not limited to pre- and postnatal care for those in low-to-middle-income countries. This technology can be modified to meet the needs of many careers and individuals to boost moods and reduce signs of depression and anxiety. Another case study exploring the power of chatbots is the work by SE Health.

Being a caregiver is an extremely selfless career, and more often than not these individuals have a tendency to take care of everyone else and forget that they have to also take care of themselves. For the caregiver, this can be overwhelming and lead to anxiety and burnout. Most cases of caregiver self-neglect are perpetuated by a lack of time and the cost of seeking treatment, posing barriers that make it almost impossible for them to receive the support and help they need.

SE Health is a Canadian nonprofit and charitable organization that services client, family, and health system needs, primarily at home and in local communities. SE Health was motivated to face these barriers head-on and make it convenient for caregivers and patients to get the support and training they need. SE Health developed a mental health chatbot named Elizzbot using X2's technology to provide support for caregiving professionals, patients, and family caregivers.

Elizzbot was repositioned to be a "helpful friend" built with a positive personality in mind the conversations are meant to be informative, empowering, and encouraging.

SE Health chose to customize an existing chatbot by taking X2AI's Tess chatbot as a starting point. Based on the knowledge of the existing staff who supports the nurses, new

content was crafted in order to make a custom version for caregivers.

Through this customizable platform and user approach, SE Health was able to gain a deeper understanding of the issues faced by caregivers and build a system that delivered personalized support to best meet each person's needs. The technology also put a focus on reinforcing the skills practiced in previous conversations through check-ins, helping to build resilience, and create unique conversations for each user [8].

An analysis of engagement patterns showed that 461 text messages on average were exchanged between Tess and the participants. The satisfaction rate was observed at an average of 88%. The chatbot measured the satisfaction rate by simply asking "Did that help?" after each chat. A positive reply was considered when the patient indicated that the intervention was helpful. Additionally, findings from a satisfaction survey found that participants felt the chatbot responded effectively with relatable topics [9].

With proof of Elizzbot's ability to support caregivers, SE Health is continuing their research and plans to strengthen the body of knowledge on psychological AI to conquer many of the traditional obstacles of mental health care many patients face. Today, Elizzbot offers emotional support at any time as a way to decrease burnout and improve prosperity to employees at SE Health, as well as signed up visitors on Elizz.com. At SE Health, these resources were evolved and scaled to support the staff, patients, and public supporters across Canada and around the world.

A new version of Elizzbot was launched as part of a grant study funded by the Baycrest Center for Aging and Brain Health Innovation (CABHI). During this study, a version of Elizzbot was offered which would read out loud the messages it would normally send. The individuals chatting with Elizzbot are also able to reply by speaking out loud. This design was found to be more appealing for older adults. The downside of this approach is the hardware cost, as a Google Home is required in order to access Elizzbot in this way [8].

Case Study: Addressing Comorbidities in Childhood Obesity, Prediabetes, and Mental Health Struggles

Many families feel the strain of childhood obesity. Along with the physical concerns, such as the potential for diabetes and hypertension, youth who are facing obesity often experience psychological symptoms. Low self-esteem and self-image can lead to disorders such as depression. It is difficult to commit to treatment goals due to missed school and work days, logistical difficulties around distance and transportation. The financial burden of co-payments and unreimbursed expenses, which could include dietitian visits and enrollment in fitness activities are barriers to care. These disorders have a 70% chance of continuing through adulthood making intervention critical in a child's long-term health.

Technological services have the potential to be utilized to help reduce these barriers, both for consumers and society as a whole. Behavioral intervention technologies (BITs) provide solutions that incorporate the benefits of technology and psychology to tackle different health needs through multiple avenues. BITs help patients in many ways, including proving accessibility*, an increased capacity for engagement, as well as a decreased cost and stigma, particularly among youth.

Artificial intelligence is one example of the most recent forms of BITs being utilized for behavior and mental health. Chatbot platforms are used for many BIT use cases as an extension of an office visit, where work done between clinician and patient can be reinforced between sessions through targeted conversations that are customized to each patient. In conjunction with a pediatric weight management program, obesity treatment, and prediabetic intervention, chatbots have the capability to build access across the tiers.

X2's chatbot Tess was customized with Nemours Children's Hospital for a weight management program for youth dealing with obesity and/or prediabetes. Nemours Children's Hospital integrated the intensive behavioral counseling recommended for weight management programs, paired with behavioral mental health approaches.

Youth with obesity symptoms enrolled in a multidisciplinary weight management program and were able to access Tess through a text-capable mobile phone number or an existing account, like FB Messenger or WhatsApp, without needing to download a separate application. The Tess integration within existing electronic health record (EHR) systems autopopulates data directly into the system without requiring extra work from the physician. Additionally, if a physician updates the EHR with mental health issues a child is experiencing, it can automatically trigger an invitation text message to be sent to the patient.

For 5 years Nemours Children's Hospital tracked which "targeted behavior" changes were most effective. For example, a young patient might have a goal of becoming a ballerina, and the strategy to reach that goal is to practice dancing for 1 h per day. This dataset of targeted behavior changes was then anonymized and shared with X2AI to develop the clinical scripts, where the physicians' responses were converted into a form that Tess could deliver. From there the scripts were read, edited by the physician and health coaches, and tested until both teams were satisfied with the outcome. The chatbot used motivational interviewing to learn which strategies the patient would want to use to achieve their goal. In the weeks after the chatbot would remind patients of these individually chosen strategies.

To measure the usability of Tess, common digital health intervention metrics were used such as satisfaction rate, engagement levels, the quantity of messages exchanged, as well as more behavior metrics including progress toward goals. Data collected based on conversations were screened three times per week during the first 2 weeks, and once per week during the following weeks. Approximately 5% of human intervention was applied for quality assurance during the early stages of this study, whereby alerts that the conversation needed support (e.g., system error or received "no" answer to a question) were sent to a person who would take over the conversation until it was corrected and Tess could resume. One such change implemented due to human intervention was the increase of conversations focused on

positive emotions, rather than a focus on negative emotions.

The results from the first cohort showed that a total of 4123 messages were exchanged between participants and Tess, resulting in 267 total conversations, which is an average of 12 conversations per patient. Adolescent patients reported experiencing positive progress toward their goals and targeted behaviors 81% of the time and patients indicated that the chats with Tess were helpful 96% of the time.

The use of an AI chatbot was found to be feasible as an adjunct to treatment, based on the adolescents in a weight management program who rated their conversations useful. The large number of messages exchanged between the chatbot and the teens, combined with the high helpfulness ratings, illustrate that this chatbot service may be an influential and engaging support tool for youth who are facing childhood obesity [10].

References

- Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M. Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: a randomized controlled trial. *JMIR Ment Health*. 2018;5:64. <https://doi.org/10.2196/mental.9782>.
- Institute for Healthcare Improvement. Triple aim for populations overview. Institute for Healthcare Improvement; 2016. <http://www.ihi.org/Topics/TripleAim/Pages/Overview.aspx>
- Lewis N. Populations, population health, and the evolution of population management: making sense of the terminology in US health care today. Institute for Healthcare Improvement; 2014. <http://www.ihi.org/communities/blogs/population-health-population-management-terminology-in-us-health-care>
- Stiefel M, Nolan K. A guide to measuring the triple aim: population health, experience of care, and per capita cost. Institute for Healthcare Improvement. 2012. <http://www.ihi.org/resources/pages/ihiwhitepapers/aguidetomeasuringtripleaim.aspx>
- Bodenheimer T, Berry-Millett R. Follow the money – controlling expenditures by improving care for patients needing costly services. *N Engl J Med*. 2009;361(16):1521–3.
- Joerin A, Rauws M, Fulmer R, Black V. Ethical artificial intelligence for digital health organizations. *Cureus*. 2020;12(3):e7202.
- Green EP, Lai Y, Pearson N, Rajasekharan S, Rauws M, Joerin A, et al. Expanding access to perinatal depression treatment in Kenya through automated psychological support: development and usability study. *JMIR Form Res*. 2020;4(10):e17895.
- Joerin A, Rauws M, Ackerman ML. Psychological artificial intelligence service, Tess: delivering on-demand support to patients and their caregivers: technical report. *Cureus*. 2019;11(1):e3972.
- Ackerman M, Virani T, Billings B. Digital mental health – innovations in consumer-driven care. *Nurs Leadersh*. 2017;30:63–72. <https://doi.org/10.12927/cjnl.2018.25384>.
- Stephens TN, Joerin A, Rauws M, Werk LN. Feasibility of pediatric obesity and prediabetes treatment support through Tess, the AI behavioral coaching chatbot. *Transl Behav Med*. 2019;9(3):440–7.



AIM in Alcohol and Drug Dependence

116

Roshan Prakash Rane, Andreas Heinz, and Kerstin Ritter

Contents

Introduction	1620
Challenges in Diagnosis and Treatment	1620
Role of Artificial Intelligence	1622
Data Sets for Machine Learning	1623
Machine Learning for Drug Dependence	1623
Challenges and Outlook	1625
Conclusion	1626
Cross-References	1626
References	1626

Abstract

Substance Use Disorders (SUD) including alcohol and drug dependence are one of the major healthcare challenges in industrialized countries. Since substance use does not always result in SUD, it is essential to

understand the complex interplay between environmental triggers (e.g., access to drugs, traumatic experiences) and biological factors (e.g., genetic predispositions, drug-induced brain alterations) that shape the development of SUD. Machine learning (ML), a subfield of artificial intelligence, has emerged recently at the forefront of medical research due to its data-driven approach and ability to model multivariate factors. It is used in SUD research to perform tasks such as disease diagnosis, predicting treatment outcomes, and identifying significant risk factors. However, the field of ML applied to SUD research is still in its infancy and faces many methodological and systemic hurdles. This chapter surveys some initial promising ML studies in SUD and discusses the main challenges for the field such as unreliable disease labels and the

R. P. Rane · A. Heinz

Department of Psychiatry and Psychotherapy, Charité – Universitätsmedizin Berlin, Berlin, Germany

K. Ritter (✉)

Department of Psychiatry and Psychotherapy, Charité – Universitätsmedizin Berlin, Berlin, Germany

Humboldt-Universität zu Berlin, Berlin, Germany

Department of Psychiatry and Psychotherapy, Berlin Institute of Health, Bernstein Center for Computational Neuroscience, Berlin, Germany
e-mail: kerstin.ritter@charite.de

sensitivity of ML algorithms toward small and heterogeneous data sets. In particular, the chapter outlines how the advances in ML coupled with an increasing availability of large and longitudinal SUD data sets could potentially transform the diagnostic process and treatment of SUD in the future.

Keywords

Machine learning · Deep learning · Substance dependence · Substance use disorder · Substance abuse · Addiction · Artificial intelligence

Introduction

The United Nations estimated that about 5.5% of the global population or 271 million people reported having consumed one or more drugs in the year 2016 [1]. In that year alone, alcohol was directly responsible for about three million deaths [2], and opioids caused a further 585 thousand deaths [1]. Direct deaths from drug use, in general, have increased over 60% from the year 2000 to 2015 [3]. Therefore, alcohol and other substance dependence is one of the major public health problems of the present times. Consuming alcohol and other harmful drugs is clinically called “substance use.” A “substance” here includes legal drugs such as alcohol, tobacco, as well as illegal drugs such as opioids (heroin, morphine, and other prescription painkillers), stimulants (cocaine, methamphetamine), hallucinogens (LSD, psilocybin mushrooms), inhalants, sedatives, depressants, hypnotics, and anxiolytics. Substance use can cause various scales of substance-induced disorders related to anxiety, mood, sleep, and sexual functioning and in rare cases can trigger psychosis [4, 5]. These effects can last anywhere from a few minutes to a few weeks based on the type of substance, the extent of intoxication, and the manner of injection. When such substance use becomes chronic and starts impairing one’s health, work, and social relationships, it is diagnosed as a psychiatric disorder called “substance dependence” or “substance use disorder” (SUD). As opposed to

substance use, substance dependence is generally characterized by an impaired ability to control the use of the substance and a persistence of use despite harm and adverse consequences [4]. For people with SUD, the motivation for substance use can begin to outweigh the drive for social and security needs such as keeping a job, saving money, and maintaining social relationships, and in adverse cases can even outweigh the need for basic physiological needs such as getting enough sleep, food, and shelter.

Challenges in Diagnosis and Treatment

Not all substance users develop a substance use disorder (SUD). Among the estimated 271 million people who consumed one or more drugs in the year 2016, only about 13% had a SUD [1]. For diagnosing SUD, medical practitioners around the world refer to the guidelines of International Classification of Diseases (ICD-11) [4] or the American Diagnostic and Statistical manual of Mental disorders (DSM-5) [6]. These diagnostic manuals suggest the clinical assessment of subjects for symptomatic addictive behaviors and effects of addiction. These include signs of inability to control substance use, substance use becoming an increasing priority in life, increased tolerance to the drug’s effects, and experiencing withdrawal symptoms [5]. While ICD-11 [4] identifies two stages of SUD development: (1) harmful pattern of substance use and (2) substance dependence, DSM-5 categorizes SUD development into three stages: (1) mild, (2) moderate, and (3) severe SUD, depending on how many of its 11 diagnostic criteria are met [6]. The current diagnostic systems, ICD and DSM, rely on clinical symptoms and signs of addiction for diagnosing and treating psychiatric disorders. They are criticized for being agnostic to the pathophysiology and the neuro-psychological mechanisms that underlie psychiatric disorders such as SUD [7, 8]. Hence, there is a shift toward developing new ways of classifying psychiatric disorders based on neurobiological measures and dimensions of psychological behavior instead [7, 9].

Over the years, researchers have identified many risk factors that contribute to the development of SUD in substance users. These risk factors can be categorized into environmental and genetic factors [10]. Prominent environmental risk factors include the ease of access to drugs, age of exposure to substance use, social facilitation, adverse life events, lack of familial support or supervision, stress, and social issues [8, 10]. Particularly, social issues such as poverty, unemployment, lack of access to housing, abusive childhood experiences, and so on form a complex web of traumatizations that contribute significantly to the development of SUD [11]. Genetic risk factors are predispositions that make an individual more prone to develop a SUD. These include neurochemical vulnerabilities in the brain [12, 13], cognitive styles [14], and other co-morbidities such as anxiety disorder and depression [15]. Such biological vulnerabilities interact dynamically with the environment, adverse life events, and the substance being consumed to ultimately determine if a substance user would develop SUD or not [8, 10, 12]. Therefore, diagnostic tools that can identify the biological vulnerabilities (or biomarkers) directly from brain imaging or even genomics would significantly improve the sensitivity of the existing diagnosis methods [7].

Active research is undergoing to identify biomarkers directly from the brain structure or function in SUD patients [16]. To study and characterize alterations in the brain, different neuroimaging techniques are employed such as structural Magnetic Resonance Imaging (sMRI), functional Magnetic Resonance Imaging (fMRI), Positron Emission Tomography (PET), and single photon emission computed tomography (SPECT). For instance, studies using PET and SPECT imaging have shown that dopamine levels in subcortical reward circuits and prefrontal cortex are associated with substance use [16]. Similarly, fMRI studies have revealed how stimulants such as cocaine and methamphetamine affects brain areas such as nucleus accumbens to produce euphoria and subsequent cravings [16]. For alcohol use disorder (AUD), a type of SUD caused by problematic consumption

of alcohol, sMRI, and fMRI studies identified numerous neurotoxic correlates in the brain such as enlarged ventricles, grey and white matter loss in frontal and reward-related brain areas, as well as altered functional connectivity in the amygdala and nucleus accumbens [17, 18]. However, these biomarkers alone are not very reliable as diagnostic tools due to their variation across the population and their overlap with other mental health problems such as depression and anxiety disorders [15]. Further advancements in neuroimaging techniques coupled with modern artificial intelligence (AI) methods could provide more reliable biomarkers from brain imaging. Section “[Machine Learning for Drug Dependence](#)” describes how current AI-based methods are being employed in SUD research to discern the effects of different biological dispositions and environmental risks, and understand their interaction with the drug itself and the employed treatment programs.

Another line of research for identifying biomarkers looks directly into genetic data for signs of SUD vulnerability. The heritability of different SUDs is estimated to range between 40% and 70% [10]. Genetics have been shown to modulate the amount of intoxicating or aversive effects caused by substance use [8]. For instance, studies have linked the susceptibility of an individual to the neurotoxic effects of chronic alcohol use with the dysfunction of the serotonin transporter (5-HTT) gene [12]. However, such genetic findings report statistical associations of risk but do not necessarily translate to diagnostics [7]. Genetic predispositions can also influence cognitive styles that can make individuals more vulnerable to SUD [9, 14]. Brain circuits that modulate certain vulnerable cognitive styles and behaviors in individuals have been associated with candidate genes such as D2 receptor, serotonin transporter (5-HTT), and the dopamine transporter (DAT) [13]. However, caution must be exercised as the risk of developing SUD due to marginalizing social factors (e.g., unemployment or restricted access to the housing market) might wrongly be affiliated to biological traits [11]. AI methods could be used with genomic data in the future

to perform multivariate analysis of genetic influences, disseminate the complex pattern of gene expressions, and identify drug targets. Thus far, the field of AI applied to genomics is still in its inception and it might require few more years for studies to emerge in this field.

Since SUD is a chronic disorder, treatment programs for SUD advocate long-term care with several follow-up sessions [19]. These programs can include medically-assisted detoxification, behavioral counseling, therapy sessions, and alleviation of withdrawal symptoms [20]. Not only can different drugs have different effects on a person's physical and mental well-being, but also the same treatment approach might show differing responses on individuals, depending on their environmental influences and their biological dispositions [20]. The disease trajectories can also be highly individualistic and include several phases of losing and regaining control over drug intake [21]. To further complicate the issue, many SUD patients usually consume more than one drug causing polysubstance effects [16]. All of these factors make the treatment of SUD not only challenging but also highly individual since no single treatment is right for everyone. To discover better treatment methods and design more individualistic treatment programs (called precision medicine), major medical research centers around the world are carrying out large-scale studies involving the gathering and analysis of data from multiple modalities such as brain imaging sessions, clinical assessments, and genomics [22, 23, 24, 25]. Furthermore, the widespread availability of smartphones and wearable electronic technology has permitted researchers to monitor behavior, cognitive-emotional states, stress reactivity and environmental exposures under real-life conditions [21]. Artificial intelligence methods can be used to analyze such large and complex data and model the effects of multiple data modalities and triggers. These AI models can then be used to predict the success of treatment programs, understand disease trajectories, and design personalized treatment programs [26, 27, 28]. AI-based studies on SUD treatment are further explored in section "Machine Learning for Drug Dependence."

Role of Artificial Intelligence

Traditionally, computer programs used for medical applications have been rule-based. This means that human programmers and experts provide explicit instructions to the computers on how to process the medical data and make decisions. In contrast, a subfield of AI called "machine learning" (ML) consists of programs that are capable of learning from example pairs of input data and the expected output to perform a specific task, without being explicitly instructed on how to perform the task. After learning the mapping between the input data and the output decision or label, the ML algorithm can be used to make predictions about new, unseen data. ML algorithms can automatically figure out the relationship between the input and the labels to solve the pre-defined task. Therefore, ML methods can be applied to complex problems consisting of several variables and data types, complex mechanisms, or unknown factors of influence, where it would not be feasible to develop rule-based methods anymore. ML is a particularly promising approach in healthcare research, since the relationship between the observed data (e.g., MRI) and the output (e.g., disease labels) can be complex, sparse, noisy, latent, and distributed across several data modalities [29]. Therefore, in the last decade, ML has emerged at the forefront of medical research in computer-assisted healthcare [29, 30].

Classically in ML, the raw data such as MRI, EEG, clinical assessments, or biochemical test reports are first processed to extract features that are deemed useful by experts for the task. For instance, for psychiatric diseases, this stage might include extraction of regional cortical thickness from sMRI or connectivity matrices from fMRI [31]. However, this step is still prone to introduce human bias into the learning process. In contrast, a new class of ML algorithm called "deep learning" is capable of learning complex and hierarchical relationships directly from raw data such as medical images or genomics [32]. Deep learning methods have surpassed human performance at tasks such as classification of images and speech recognition [32]. In medical imaging applications including skin cancer

diagnosis and brain tumor segmentation, they currently hold state-of-the-art performance [30]. Some initial promising deep learning studies based on neuroimaging data have also been published for psychiatric diseases such as Alzheimer's disease and schizophrenia [45].

Section “[Data Sets for Machine Learning](#)” highlights the significance of data sets to ML research by introducing some of the popular data sets in SUD. Next, section “[Machine Learning for Drug Dependence](#)” surveys the preliminary ML-based studies in substance dependence and discusses their ongoing contributions to the field of SUD research. Finally, section “[Challenges and Outlook](#)” discusses the current challenges that must be overcome to build reliable ML solutions for SUD diagnosis and treatment. The section also presents the promises of ML and its possible future applications in substance dependence.

Data Sets for Machine Learning

As discussed in section “[Challenges in Diagnosis and Treatment](#),” SUD is a chronic disorder that is characterized by the complex interactions between environmental and biological risk factors. To congregate these different influencing factors and understand their complex interactions, a number of multi-site studies have been carried out, e.g., IMAGEN [22], UK Biobank [23], NCANDA [24], and ENIGMA [25]. These studies acquire large data sets consisting of hundreds or thousands of subjects and for each subject, multiple domains of data are collected depending on the study design. Such data modalities, for example, include brain imaging, genomics, clinical assessments, psychiatric tests, self-reports, and demographic data [22, 23, 24, 25]. To better understand the dynamics of losing and gaining control over drug intake in SUD, longitudinal and – if possible – real-time data (e.g., acquired via smartphones) are essential [21]. Such large-scale data collection initiatives are the key to future research in SUD and other major psychiatric disorders. These data sets not only facilitate large-scale

statistical analyses but also enable the development of novel methodologies based on ML and particularly deep learning.

Machine Learning for Drug Dependence

In this section, existing ML studies in SUD are surveyed and their implications to SUD research are discussed.

ML methods promise a data-driven approach to diagnosing SUD, generating individualized treatment recommendations, and predicting treatment outcomes. Since its inception, a number of different ML methods have been invented [32] and some of them have been applied to SUD. These include Logistic Regression, Random Forest, Decision Trees, and Support Vector Machines (SVM) that are applied to hand-crafted features (e.g., regional brain volumes extracted from sMRI) [26, 27, 28, 31, 33, 34, 35, 38, 39, 40, 41], as well as modern deep learning methods such as Fully-Connected Neural networks (FCN), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN) [42, 43, 44] which can be directly applied to raw data (e.g. sMRI). These ML models are trained either on a single data domain (e.g., just sMRI) or a combination of multiple domains (e.g., demographic data in combination with cognitive testing and sMRI) [26, 33, 36, 37, 38]. ML models have been trained to perform several tasks such as separating patients with SUD from healthy controls and mild substance users [26, 31], predicting the development of SUD in the future [26, 37], predicting treatment success [27, 28], and identifying the most informative risk factors [37, 38, 40]. Figure 1 demonstrates how ML is used to solve some of the challenges in SUD diagnosis and treatment.

So far, most ML studies focus on AUD, since alcohol dependence is one of the prominent SUDs and large data sets are available for it [22, 24]. A common application is to infer the influence of different risk factors and data domains for the onset of AUD. Squeglia et al. [37] trained ML models on the data of adolescents in the age group of 12 to 14 to predict the development of moderate to high alcohol use at the age of 18. They trained on a combination of demographic,

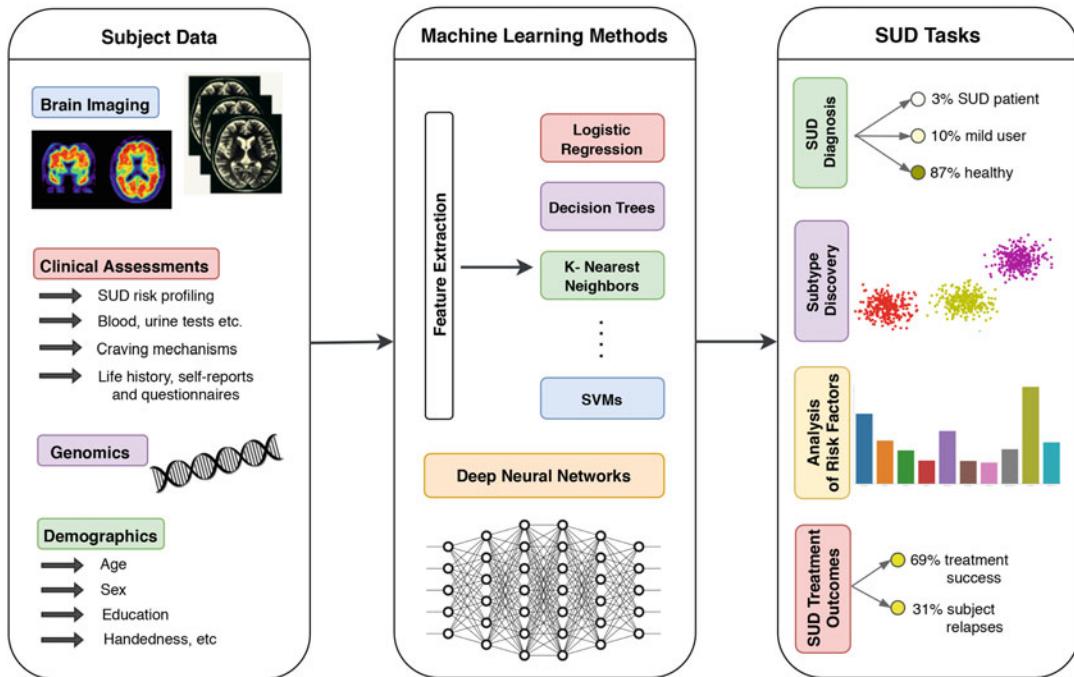


Fig. 1 An illustration of how Machine Learning (ML) is used in Substance Use Disorder (SUD) research. The first column lists some of the commonly used data modalities including brain imaging and clinical assessments, and the second column lists some of the classical ML and deep learning methods. The models can be trained on a single

data domain (e.g., brain imaging) or on multiple data domains (e.g., clinical, demographic and imaging data). The third column lists some of the common applications in SUD including diagnosis, discovering neurobiological disease subtypes, determining the influence of different risk factors, and studying treatment outcomes

neurocognitive, and neuroimaging data and identified 34 factors which best predicted future development of alcohol abuse. These included worse executive functioning, behavioral factors, thinner cortices, and reduced activations in frontal and temporal areas of the brain. A multi-data-domain study showed that brain connectivity matrices obtained from resting-state fMRI were best at predicting alcohol use severity, compared to other data domains such as demographic data, sMRI, or task-based fMRI [36]. Another study based on resting-state fMRI data reported that the so-called reward network and the executive control network in the brain were most informative for diagnosing AUD [35]. The largest classical ML study so far has been performed on the IMAGEN data set by Whelan et al. [26] who report an AUC (Area under the receiver-operator characteristic curve) of 0.90–0.96 for the separation of 14-year old binge drinkers of alcohol and

14-year old controls and an AUC of 0.75 for predicting binge drinking at 16 years. Here, a combination of history, social experiences (e.g., romantic relationships), and brain features was shown to be most predictive, supporting the hypothesis that multiple causal factors shape later alcohol use. Guggenmos et al. showed that an experienced radiologist scored 66% at diagnosing AUD from sMRI data, while their ML method achieved a balanced accuracy of 74% on the same sMRI data [31]. These results imply statistically significant neurophysiological differences among AUD and healthy controls. Few recent studies have also applied deep learning models to identify subjects with AUD [43, 44]. They trained Convolutional neural network (CNN) models on 2-dimensional slices of sMRI data and reported accuracies of up to 97% in discriminating patients with abstinent long-term chronic AUD and healthy controls.

ML studies have also been performed on subjects with other substance dependencies. A multivariate ML study found that different clinical assessment factors are useful for identifying subjects with heroin dependence as compared to subjects with amphetamine dependence, suggesting that predictive markers can be very substance-specific [39]. In a small-sample study on subjects with cocaine use disorder, researchers showed that phenotypes of high impulsivity were most predictive of developing cocaine dependence [40]. For cannabis dependence, a study based on the IMAGEN data set identified a risk profile containing psychosocial and sex-specific brain prognostic markers, which were likely to precede and influence cannabis initiation [41]. This study also underlines the importance of considering sex as an influencing factor in SUD. A very recent ML study found that the main prognostic features for development of SUD changed with age [38]. While psychological dysregulation best predicted SUD during childhood and early adolescence, social and interpersonal interaction problems were most predictive features in late adolescence and thereafter. Another line of research is to use social media with deep learning methods to identify the risk of alcohol, tobacco, and drug use [42]. While acknowledging the user privacy issues that must be addressed, they show that this could be a potentially new way of estimating substance use risks [42].

ML methods have also been applied to develop treatment programs for SUD and predict treatment outcomes. A recent study used decision trees to predict whether an individual with AUD seeks treatment and found that the total drinks consumed in lifetime and not having depression or other substance disorders were the strongest predictors of seeking treatment [28]. Similarly, in [27] the performance of different classical ML have been compared in order to predict SUD treatment. Another direction of research involves using smartphones or personal digital assistants to monitor psychological profiles and detect drug use in real-time [21]. ML methods can be used along with these ambulatory data to automatically perform interventions, recommend treatment programs, and study the outcome of treatments [47].

Challenges and Outlook

The field of ML applied to healthcare is still in its infancy [30]. The previous section surveyed preliminary ML-based research in SUD that generally used small sample sizes and have not yet undergone rigorous and independent replication tests. Their findings have also not yet been critically assessed in a clinical setting. Therefore, one can conclude that the true potential of ML (and especially deep learning) are still under-exploration in SUD research [30]. Currently, there are several hurdles for implementing reliable ML methods for SUD. The foremost challenge is the ambiguity of clinical labels in SUD and their high overlap with other psychiatric diseases [9]. Current psychiatric labels defined by ICD-11 or DSM-5 were designed to characterize observable symptoms of addiction and identify problematic behaviors. They are often criticized for having a multifold biopsychological basis that overlap between multiple mental disorders [5, 7]. Therefore, existing psychiatric labels are not reliable enough for designing a learning framework for ML models due to their heterogeneity in clinical presentation and reliance on clinical symptoms rather than neurobiological substrates. Another major hurdle in ML is the need for reliable, large-scale, and open-access data sets for developing methods and benchmarking progress in the field. One of the primary reasons for the success of ML in applications such as computer vision and natural language processing is the availability of high-quality data sets with large sample sizes (for example, a popular image classification data set, ImageNet has 14 million images) [32]. Large-scale data collection and harmonization initiatives have only just begun in SUD research [22, 23, 24, 25]. Furthermore, data sets in medical domains such as neuroimaging can have high variability based on factors such as the imaging site, the imaging protocol, and the instruments used [22]. This heterogeneity across data sets (and sometimes within a data set collected in multiple sites) poses a challenge for developing generalizable ML solutions. In order to overcome this, data collection studies are required to not only harmonize their acquisition protocols, tools,

and pre-processing pipelines across several imaging sites but are also required to collect a large number of samples per site that is representative of the population distribution. The chronic nature of SUD further exacerbates the problem. It mandates the collection of longitudinal data of subjects over several years so that one can reliably capture the dynamics of relapse and triggers, understand the complex trajectories in the development of drug dependence, and quantify the success of treatment approaches [21]. ML also struggles with several methodological challenges in applications such as medical imaging. For instance, neuroimaging data such as sMRI or fMRI have usually a low signal-to-noise ratio and thus current ML methods require large sample sizes and computational capacities to learn the underlying useful signal for the task [30]. Prevailing ML methods like deep learning have an algorithmic bias toward natural images. They are also criticized for being a black-box as their decisions are currently hard to interpret, although adhoc visualization techniques exist [45, 46]. However, these hurdles can mostly be attributed to the field of ML and deep learning still being in its initial stages of research in medical applications [30].

Conclusion

Substance use disorder (SUD) is a chronic mental health disorder that is affected by a complex interplay of several environmental triggers (e.g., access to drugs, traumatic experiences) and biological factors (e.g., genetic predispositions). Machine learning (ML) and artificial intelligence, in general, have emerged as promising tools in the field. Despite some methodological hindrances, ML has proven to be capable of capturing associations from multiple influencing factors and model the complex interactions in psychiatric disorders such as SUD [26, 37]. Therefore, as more and more large-scale and standardized data sets become available for SUD, ML methods would exceedingly contribute to SUD research in novel and beneficial ways. Due to its data-driven approach, ML can help in the discovery of neurobiologically valid psychiatric labels, in

contrast to the existing labels. In the future, methods based on ML could lead to the development of new methods of diagnosing SUD [29] and understanding the relationship between SUD and other mental disorders. Furthermore, ML could be the key to enable precision medicine by recommending personalized patient treatments, predicting treatment outcomes [28, 27], and guiding the development of new therapies [47].

Cross-References

- AIM and Explainable Methods in Medical Imaging and Diagnostics

References

1. World Drug Report. 35 million people worldwide suffer from drug use disorders while only 1 in 7 people receive treatment. 2019. <http://www.unodc.org>. Accessed 15 Sept 2020.
2. WHO. Global status report on alcohol and health 2018. 2018. p. xvi. Accessed 15 Sept 2020.
3. World Drug Report. Prelaunch. 2018. <http://www.unodc.org/wdr2018/prelaunch/>. Accessed 15 Sept 2020.
4. World Health Organization. International classification of diseases for mortality and morbidity statistics (11th Revision). 2018. <https://icd.who.int/browse11/l-m/en>. Accessed 15 Sept 2020.
5. Saunders JB. Substance use and addictive disorders in DSM-5 and ICD 10 and the draft ICD 11. *Curr Opin Psychiatry*. 2017;30(4):227–37.
6. American Psychiatric Association, American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-5. United States; 2013.
7. Insel TR. The NIMH research domain criteria (rdc) project: precision medicine for psychiatry [Internet]. *Am J Psychiatr*. 2014;171:395–7. Available from: <http://www.nimh.nih.gov/research-priorities/rdoc/index.shtml>
8. Heinz AJ, Beck A, Meyer-Lindenberg A, Sterzer P, Heinz A. Cognitive and neurobiological mechanisms of alcohol-related aggression. *Nat Rev Neurosci*. 2011;12(7):400–13. <https://doi.org/10.1038/nrn3042>. PMID: 21633380
9. Heinz A. A new understanding of mental disorders: computational models for dimensional psychiatry. Cambridge, MA: MIT Press; 2017.
10. Kendler KS, Chen X, Dick D, Maes H, Gillespie N, Neale MC, et al. Recent advances in the genetic epidemiology and molecular genetics of substance use

- disorders. *Nat Neurosci.* 2012;15:181–9. Available from: <https://pubmed.ncbi.nlm.nih.gov/22281715/>
11. Heinz A, Zhao X, Liu S. Implications of the association of social exclusion with mental health. *JAMA Psychiatry.* 2020;77(2):113–4. <https://doi.org/10.1001/jamapsychiatry.2019.3009>.
 12. Heinz A, Jones DW, Mazzanti C, Goldman D, Ragan P, Hommer D, Linnoila M, Weinberger DR. A relationship between serotonin transporter genotype and in vivo protein expression and alcohol neurotoxicity. *Biol Psychiatry.* 2000;47(7):643–9.
 13. Belcher AM, Volkow ND, Moeller FG, Ferré S. Personality traits and vulnerability or resilience to substance use disorders. *Trends Cogn Sci.* 2014;18(4):211–7. <https://doi.org/10.1016/j.tics.2014.01.010>.
 14. Robbins T, Everitt B. Drug addiction: bad habits add up. *Nature.* 1999;398:567–70. <https://doi.org/10.1038/19208>.
 15. Kotov R, et al. Linking “big” personality traits to anxiety, depressive, and substance use disorders: a meta-analysis. *Psychol Bull.* 2010;136:768–821. [PubMed: 20804236]
 16. Fowler JS, Volkow ND, Kassed CA, Chang L. Imaging the addicted human brain. *Sci Pract Perspect.* 2007;3(2):4–16. <https://doi.org/10.1151/spp07324>.
 17. Bühlér M, Mann K. Alcohol and the human brain: a systematic review of different neuroimaging methods. *Alcohol Clin Exp Res.* 2011;35(10):17711793.
 18. Zahr NM, Pfefferbaum A. Alcohol’s effects on the brain: neuroimaging results in humans and animal models. *Alcohol Res Curr Rev.* 2017;38(2):183–206.
 19. National Institute on Drug Abuse. Treatment approaches for drug addiction DrugFacts. 2019. <https://www.drugabuse.gov/publications/drugfacts/treatment-approaches-drug-addiction>. Retrieved 15 Sept 2020.
 20. Walters ST, Rotgers F, editors. Treating substance abuse: theory and technique. New York: Guilford Press; 2011.
 21. Heinz A, Kiefer F, Smolka MN, Endrass T, Beste C, Beck A, et al. Addiction Research Consortium: losing and regaining control over drug intake (ReCoDe) – from trajectories to mechanisms and interventions. *Addict Biol [Internet].* 2020;25(2) <https://doi.org/10.1111/adb.12866>.
 22. Mascarell Maričić L, Walter H, Rosenthal A, Ripke S, Quinlan EB, Banaschewski T, et al.. The IMAGEN study: a decade of imaging genetics in adolescents [Internet]. *Mol Psychiatry.* 2020. Springer Nature, p. 1–24. <https://doi.org/10.1038/s41380-020-0822-5>
 23. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* 2015;12(3):e1001779.
 24. Brown SA, Brumback T, Tomlinson K, Cummins K, Thompson WK, Nagel BJ, De Bellis MD, Hooper SR, Clark DB, Chung T, Hasler BP. The National Consortium on Alcohol and NeuroDevelopment in Adolescence (NCANDA): a multisite study of adolescent development and substance use. *J Stud Alcohol Drugs.* 2015;76(6):895–908.
 25. Thompson PM, Stein JL, Medland SE, Hibar DP, Vasquez AA, Renteria ME, Toro R, Jahanshad N, Schumann G, Franke B, Wright MJ. The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain Imaging Behav.* 2014;8(2):153–82.
 26. Whelan R, Watts R, Orr CA, Althoff RR, Artiges E, Banaschewski T, Barker GJ, Bokde AL, Büchel C, Carvalho FM, Conrod PJ. Neuropsychosocial profiles of current and future adolescent alcohol misusers. *Nature.* 2014;512(7513):185–9.
 27. Acion L, Kelmansky D, van der Laan M, Sahker E, Jones D, Arndt S. Use of a machine learning framework to predict substance use disorder treatment success. *PLoS One.* 2017;12(4):e0175383.
 28. Lee MR, Sankar V, Hammer A, Kennedy WG, Barb JJ, McQueen PG, Leggio L. Using machine learning to classify individuals with alcohol use disorder based on treatment seeking status. *EClinical Med.* 2019;12:70–8.
 29. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med.* 2019;380(14):1347–58.
 30. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, Ferrero E, Agapow PM, Zietz M, Hoffman MM, Xie W. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface.* 2018;15(141):20170387.
 31. Guggenmos M, Scheel M, Sekutowicz M, Garbusow M, Sebold M, Sommer C, Charlet K, Beck A, Wittchen H-U, Zimmermann U, et al. Decoding diagnosis and lifetime consumption in alcohol dependence from grey-matter pattern information. *Acta Psychiatr Scand.* 2018;137:252–62.
 32. Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep learning, vol. 1. Cambridge: MIT Press; 2016.
 33. Guggenmos M, Schmack K, Veer IM, Lett T, Sekutowicz M, Sebold M, Garbusow M, Sommer C, Wittchen H-U, Zimmermann US, et al. A multimodal neuroimaging classifier for alcohol dependence. *Sci Rep.* 2020;10:1–12.
 34. Seo S, Mohr J, Beck A, Wüstenberg T, Heinz A, Obermayer K. Predicting the future relapse of alcohol-dependent patients from structural and functional brain images. *Addict Biol.* 2015;20:1042–55.
 35. Zhu X, Du X, Kerich M, Lohoff FW, Momenan R. Random forest based classification of alcohol dependence patients and healthy controls using resting state MRI. *Neurosci Lett.* 2018;676:27–33.
 36. Fede SJ, Grodin EN, Dean SF, Diazgranados N, Momenan R. Resting state connectivity best predicts alcohol use severity in moderate to heavy alcohol users. *NeuroImage Clin.* 2019;22:101782.
 37. Squeglia LM, Ball TM, Jacobus J, Brumback T, McKenna BS, Nguyen-Louie TT, Sorg SF, Paulus MP, Tapert SF. Neural predictors of initiating alcohol use during adolescence. *Am J Psychiatr.* 2017;174:172–85.

38. Jing Y, Hu Z, Fan P, Xue Y, Wang L, Tarter RE, Kirisci L, Wang J, Vanyukov M, Xie XQ. Analysis of substance use and its outcomes by machine learning I. Childhood evaluation of liability to substance use disorder. *Drug Alcohol Depend.* 2020;206:107605. <https://doi.org/10.1016/j.drugalcdep.2019.107605>. Epub 2019 Oct 22. PMID: 31839402; PMCID: PMC6980708.
39. Ahn WY, Vassileva J. Machine-learning identifies substance-specific behavioral markers for opiate and stimulant dependence. *Drug Alcohol Depend.* 2016;161:247–57.
40. Ahn WY, Ramesh D, Moeller FG, Vassileva J. Utility of machine-learning approaches to identify behavioral markers for substance use disorders: impulsivity dimensions as predictors of current cocaine dependence. *Front Psychiatry.* 2016;7:34. <https://doi.org/10.3389/fpsyg.2016.00034>.
41. Spechler PA, Allgaier N, Chaarani B, Whelan R, Watts R, Orr C, Albaugh MD, D'Alberto N, Higgins ST, Hudson KE, et al. The initiation of cannabis use in adolescence is predicted by sex-specific psychosocial and neurobiological features. *Eur J Neurosci.* 2019;50: 2346–56.
42. Hassanpour S, Tomita N, DeLise T, Crosier B, Marsch LA. Identifying substance use risk based on deep neural networks and Instagram social media data. *Neuropsychopharmacology.* 2019;44(3):487–94. <https://doi.org/10.1038/s41386-018-0247-x>.
43. Wang S-H, Xie S, Chen X, Guttery DS, Tang C, Sun J, Zhang Y-D. Alcoholism identification based on an alexnet transfer learning model. *Front Psychiatry.* 2019;10:205.
44. Wang S-H, Muhammad K, Hong J, Sangaiah AK, Zhang Y-D. Alcoholism identification via convolutional neural network based on parametric ReLU, dropout, and batch normalization. *Neural Comput Appl.* 2020;32:665–80.
45. Böhle M, Eitel F, Weygandt M, Ritter K. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. *Front Aging Neurosci.* 2019;11:194.
46. Eitel F, Soehler E, Bellmann-Strobl J, Brandt AU, Ruprecht K, Giess RM, Kuchling J, Asseyer S, Weygandt M, Haynes JD, Scheel M. Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation. *NeuroImage: Clinical.* 2019;24:102003.
47. Bertz JW, Epstein DH, Preston KL. Combining ecological momentary assessment with objective, ambulatory measures of behavior and physiology in substance-use research. *Addict Behav.* 2018;83:5–17. <https://doi.org/10.1016/j.addbeh.2017.11.027>.



Artificial Intelligence in Medicine and PTSD

117

Victor Trouset and Thomas Lefèvre

Contents

PTSD: A Complex Clinical Disorder	1630
PTSD: A Psychological Disorder that May Appear After Exposure	
to a Traumatic Event	1630
PTSD: A Disorder that Is Currently Difficult to Explain and Predict	1630
The Evolution of PTSD Over Time May Be Complex and Chronic, and	
Associated with Physical and Mental Disorders of a Different Nature	1631
Improving Screening for and Diagnosis of PTSD	1632
Present Advances: AI and PTSD	1633
AI and PTSD Prediction. AI for Clinical Practice and Practitioners	1633
AI, Characterization and Diagnosis of PTSD. AI for Basic Research	1635
Potential Trends and Future Challenges	1638
References	1639

Abstract

Post-traumatic stress disorder (PTSD) is a relatively recently defined mental disorder which is diagnosed with the aid of a set of criteria specified by the DSM. PTSD results from a traumatic event and involves clusters of symptoms which together disrupt the psychological, physical, and social functioning of the person.

Although about 70% of people will be exposed at least once in their lifetime to a traumatic event, and that about 4% of them will develop PTSD (chronic or not), screening and diagnosis of the disorder is often seen to be primarily a matter for specialist psychiatrists. Not only is there a lack of awareness of PTSD among health professionals and insufficient training in identifying the disorder, there is also a shortage of relevant diagnostic tools in routine clinical practice. The identified determinants and risk factors for PTSD are of little use in either explaining the disorder satisfactorily or identifying it effectively. This has an effect on primary and secondary prevention, while few treatments, either chemical or non-chemical, have been shown to be effective to date. It is therefore worth asking whether artificial

V. Trouset (✉) · T. Lefèvre
IRIS Institut de Recherche Interdisciplinaire sur les enjeux Sociaux, UMR8156 CNRS – U997 Inserm – EHESS – Université Sorbonne Paris Nord, Paris, France

Department of Forensic and Social Medicine, AP-HP, Jean Verdier Hospital, Bondy, France
e-mail: victor.trouset@aphp.fr;
thomas.lefeuvre@univ-paris13.fr

Intelligence might make a contribution in any of the dimensions mentioned above: prediction, diagnosis, etiology, or cure. Could AI improve understanding of the disorder, compared to the more traditional techniques of clinical research and epidemiology? And might it be an important advance in the management of PTSD? Few existing studies on AI and PTSD, and it is difficult to make comparisons between those that do exist.

Keywords

Post-traumatic stress disorder · Traumatic event · Prediction · Diagnosis · Risk factors · Predictors · Machine learning · Artificial intelligence · Neuroimaging data

PTSD: A Complex Clinical Disorder

PTSD: A Psychological Disorder that May Appear After Exposure to a Traumatic Event

PTSD is a psychological disorder that usually occurs within the first 3 months of exposure to a traumatic event. This syndrome can also appear sometime after the traumatic event, with a delay of a few months or even years. Exposure to a traumatic event – a sudden situation that leads to a person's mental defense mechanisms being overwhelmed – is one of the conditions needed for a diagnosis of PTSD. Almost 70% of people will experience in their lifetime at least one traumatic event among a set of 29 traumatic events designed by the WHO World Mental Health (WMH) Surveys and required for a PTSD diagnosis [1]. Traumatic events may be intentional, as, for example, in interpersonal violence, (fighting, torture, mutilation, physical violence, sexual assault, murder of a loved one), or they may be unintentional, as in road accidents, natural disasters and serious illness (Table 1). Not all those who experience traumatic events go on to suffer from PTSD, however, since among adult individuals exposed to at least one traumatic event, lifetime prevalence of PTSD stands at only about 4% [1]. The identification and

nosological diagnosis of PTSD is recent. Diagnostic criteria have evolved since the first appearance of PTSD in the third version of the Diagnostic and Statistical Manual of Mental Disorders (DSM) in 1980, while in its current version (DSM-5) the presence of seven other criteria are needed for a diagnosis of PTSD (Table 2). PTSD is a multi-faceted, primarily clinical diagnosis that is likely to include numerous combinations of more or less specific symptoms [2]. The variability in the clinical expression of PTSD explains part of the difficulty in using classifications such as DSM-5 in screening for this disorder, since the aim of the DSM is to provide standardized diagnoses for comparability and reproducibility in research and is not designed for clinical use and diagnosis.

Screening for psychological symptoms that might indicate PTSD is essential for patient referral and for preventing the development of disabling complications, so healthcare professionals need to be trained in screening in order to provide valid diagnoses. Current diagnostic tools are accessible to all doctors, but their use requires specific experience in psycho-traumatology, while clinical interviews or self-report measures for PTSD are time-consuming and not necessarily well-suited to routine clinical practice.

Specify whether with dissociative symptoms (dissociative subtype of PTSD). In addition to meet the full criteria for PTSD, the individual persistently or repeatedly experiences one of two dissociative symptoms following: depersonalization or derealization.

PTSD: A Disorder that Is Currently Difficult to Explain and Predict

There have been numerous studies on the predictive factors and determinants of PTSD since the disorder first appeared in the nosography. This research is essential for understanding and screening for the disorder and referring individuals to specific care pathways. Certain determinants explain the disparity that has been noted between the frequency of exposure to a traumatic event and the frequency of PTSD. Although several clinical

Table 1 Classification of different types of potentially traumatic events by intentional or non-intentional origin

Non-intentional events	
1.	Accident, serious injury or illness
	Natural disaster
	Man-made disaster
	Toxic chemical exposure
	Life-threatening motor vehicle accident
	Other life-threatening accident
	Life-threatening illness
	Life-threatening illness or injury of child
Interpersonal violence	
2.	Experience of sexual violence
	Raped
	Sexually assaulted
	Stalked (sexual harassment)
	Any other private event
3.	Experience of physical violence
	Beaten up by spouse or intimate partner
	Beaten up by someone else
	Beaten up by caregiver (childhood physical abuse)
	Having been witnessed intimate partner violence or domestic violence
	Mugged or threatened with a weapon
4.	Exposure to organized violence
	Relief worker in war zone
	Civilian in war zone
	Civilian in region of terror
	Refugee
	Kidnapped
5.	Participation in organized violence
	Witnessed death or serious injury, or discovered dead body
	Witnessed atrocities
	Combat experience
	Unintentionally caused death or serious injury
	Purposely injured, tortured or killed someone
6.	Other
	Unexpected death of loved one (homicide, accident, suicide, fatal illness)
	Other traumatic event of loved one (kidnapped, tortured, or raped)
	Any other traumatic or life-threatening event

determinants have been proposed, studies have shown that the results on their identification and their respective weight in the occurrence of PTSD are variable and sometimes contradictory. One study found that pre-traumatic factors, such as gender, with women being more likely to suffer

PTSD (OR = 1.98, IC95% 1.76–2.22) [3], may be more predictive than other factors, while for other researchers, a peritraumatic factor (peritraumatic dissociation) and a post-traumatic factor (lack of social support) were more predictive of PTSD, albeit with a contribution that was modest (mean correlation coefficient r : 0.35 and 0.40, respectively) and heterogeneous (r = 0.14 to 0.94 and r = −0.02 to 0.54, respectively) [4, 5]. Conversely, a 2019 meta-analysis that included more types of traumatic events and referred to several parameters to evaluate the strength of the association showed that the existence of physical vulnerability or a family history of psychiatric disorder are the best predictors of PTSD [6]. Other types of determinants that facilitate the diagnosis of PTSD, such as potential biomarkers, including neuroendocrine, genetic, neuroanatomical, and neurofunctional abnormalities, have also been identified [7]. There is still some speculation about the role of these biological correlates in the occurrence of PTSD (they lack robustness and validity for use in current practice) but they do contribute to a better understanding of the biopsychological mechanism of PTSD and have specific implications for the future: specifying PTSD via these biomarkers may make it possible to achieve a more accurate diagnosis and offer more targeted therapies.

Due to the high number of potentially traumatic events, any kind of doctor, whether in a hospital or in primary care, may well encounter a patient with symptoms of PTSD. Professionals' lack of awareness of psychotraumatic stress and the absence of diagnostic tools that can easily be used by any doctor lead to under-diagnosis of PTSD.

The Evolution of PTSD Over Time May Be Complex and Chronic, and Associated with Physical and Mental Disorders of a Different Nature

The symptoms of PTSD, particularly certain symptoms, may evolve over time. In half of cases, PTSD symptoms disappear within the first 3 months. Conversely, PTSD may become chronic with the persistence of symptoms for

Table 2 DSM-5 diagnostic criteria for PTSD [41]

A. The person experienced, witnessed, or was confronted with one of the following traumatic event(s): actual or threatened death, serious injury or sexual violence
B. 1 (or more) of 5 reexperiencing symptoms:
Intrusive, repetitive and disabling recollections
Repetitive distressing nightmares
Acting/feeling as though event were recurring (dissociative flashbacks)
Psychological distress when exposed to reminders of the traumatic event
Physiological hyperreactivity when exposed to reminders of the traumatic event
C. 1 (or more) of 2 avoidance symptoms:
Efforts to avoid recollections, thoughts or feelings about the traumatic event
Efforts to avoid external reminders (people, places, talking, activities, objects or situations) that trigger recollections, thoughts or feelings associated with the traumatic event
D. 2 (or more) of 7 negative alterations in cognition and mood:
Inability to recall important aspects of the traumatic event (dissociative amnesia)
Negative beliefs or expectations about oneself, other people or the world
Cognitive distortions about the traumatic event leading to blame oneself or other people
Persistent negative emotional state (e.g., fear, horror, anger, guilt or shame)
Diminished interest or participation in significant activities
Detachment from others
Inability to experience positive emotions or feelings
E. 2 (or more) of 6 hyperarousal symptoms:
Irritability or angry outbursts with verbal or physical assaults towards people or objects
Reckless or self-defeating behavior
Hypervigilance
Exaggerated startle response
Concentration disorders
Sleep disorders
F. Duration of the symptoms is at least 1 month
G. Requires clinically significant distress or functional impairment (e.g., social, occupational)
H. Exclude differential diagnosis

12 months or for several years [8]. The medium to long-term evolution of PTSD is characterized by the frequent co-occurrence of at least one other mental disorder (e.g., depression, anxiety disorders, substance use disorders) [9] and an increased risk of suicide (RR = 2.7, 95% CI 1.3–5.5) [10]. It is also associated with a negative perception of one's health, the appearance of numerous unexplained physical complaints, and an impact on social life (e.g., social isolation) and work (e.g., losing one's job) [11]. The physical impact may be even greater, and related to the different neurobiological reactions to stress, for example, an increased risk of cardiovascular events. The impact of PTSD on a person's life thus seems to be all-embracing and profound. At the very least, it can be part of a complex and disabling clinical and biographical picture, whether as a cause,

consequence or comorbidity. All these complications contribute to masking or aggravating the symptoms of PTSD and can lead to delayed diagnosis, inappropriate management of the disorder and its comorbidities, and ultimately, to a burden on the healthcare system.

Improving Screening for and Diagnosis of PTSD

There are therefore strong arguments for improving screening and diagnosis of PTSD in order to ensure that treatment of people who may have this disorder is conducted early, or at the very least, delayed as little as possible. To date, none of the current treatments available are highly or even systematically effective. However, the earlier the

diagnosis is made, or even anticipated (through secondary prevention following exposure to trauma), the easier it will be to circumvent the effects of PTSD on the social and individual functioning of the person.

In general terms, the issues related to PTSD, to date, can be categorized as follows:

- (i) Diagnosis, screening (particularly by non-specialists) and patient referral within the healthcare system, which are all problematic because of the high cost and the shortage of specialists able to deal with this widespread problem and because current diagnostic tools (DSM and evaluation scales) are essentially oriented towards epidemiological research.
- (ii) Primary and secondary prevention: ideally, following exposure to trauma, occurrence of PTSD should be avoided and if it does occur, its duration and its impact on the social and individual functioning of the person should be limited.
- (iii) Therapies.
- (iv) Research: better understanding, better definition, and better characterization of PTSD, its different forms and expressions, and its evolution over time; research into causes; research into associations with other psychological and physical diagnoses in order to improve understanding of the network of multiple causes and consequences which PTSD is part of.

The purpose of this chapter is to examine the potential value of AI in relation to PTSD from two complementary perspectives:

- (i) In relation to existing diagnostic, explanatory and predictive methods
- (ii) For possible new/original uses, unique to AI or enabled by it

Present Advances: AI and PTSD

The use of AI in PTSD is still marginal when compared to its use in other psychiatric disorders, although there is recent and growing interest in exploring new predictive methods based on

it. There are applications in two main areas: prediction and diagnosis. Data sources that can be used in models are varied and of different types, including clinical, biological, genomic and neuroimaging data.

AI and PTSD Prediction. AI for Clinical Practice and Practitioners

In this section we discuss the potential contribution of AI in predicting (i) a diagnosis of PTSD, and (ii) response to treatment. With regard to diagnosis, it is accepted that a diagnosis of PTSD can only be made 1 month after exposure to a traumatic event. A central question may therefore be whether during this period one is able to identify and anticipate that PTSD may occur within 1 month – or more – of exposure to trauma. With regard to treatment, as in all 4P medicine, the aim is to predict and personalize the care and treatment of individuals. In mental health, perhaps more than in other fields, where therapeutic indications may be porous, depending on which diagnosis is made, and response varies greatly from one person to another, it is valuable to know if AI can help practitioners to estimate the potential effectiveness of a potential treatment and to be more precise and so save time. Unsurprisingly, the majority of studies to date have been secondary, retrospective analyses of cohort epidemiological data or cohorts reconstituted from hospital data.

Early Prediction of PTSD

Identifying subjects at risk of developing PTSD after a traumatic event is still a challenge for standard determinant approaches. Early prediction of PTSD by AI on initial data alone and before the disorder is diagnosed conclusively (1 month later) would enable more rapid intervention and hopefully avoid the emergence of chronic, invasive, and disabling symptoms.

Few studies have used AI for early prediction of PTSD, the majority having been conducted in hospital, more specifically, in emergency departments.

Several researchers have used various types of clinical, biological and psychological data collected in emergency departments for the early

prediction of PTSD [12–16]. Initial data were collected immediately “at the patient’s bedside” in the emergency room and sometimes supplemented by other data, particularly psychological or biological data, collected within 10 days of the traumatic event [14–16]. The clinical data included physiological variables usually measured in the emergency department (e.g., heart rate and blood pressure) and the biological data collected in the emergency department or at discharge included standard biological variables or salivary, urinary and blood neuroendocrine variables involved in the response to stress. Prediction using only data collected in the emergency department showed a prediction capacity ranging from accurate to good ($AUC = 75$ to 85%) regardless of the type of data used in the prediction model. The best predictive performance was obtained by combining both clinical and biological data, in contrast to the predictive capacity of biological data alone ($AUC = 72\%$ in Schultebraucks 2020; $AUC = 67\%$ in Galatzer-Levy 2017). Predictive performance on data collected in two stages was similar ($AUC = 82\%$ in Galatzer-Levy 2014) or even slightly better ($AUC = 88\%$ in Galatzer-Levy 2017), but insufficient if based on biological data alone ($AUC = 66\%$ in Galatzer-Levy 2017).

The use of biological data collected in emergency department seems appropriate as it can be collected from technical platforms available to doctors, but its use alone does not lead automatically to improvement in predictive performance and early detection. Biological data alone is not enough to predict PTSD, and this is corroborated by the fact that no biological determinants have been clearly identified as predictive of PTSD. The fact that predictive models performed best when psychological variables were taken into account underlines the pre-eminence of clinical data in the early prediction of PTSD. This finding is consistent with the definition of PTSD based on symptoms; the use of questionnaires in an emergency context, however, seems difficult to transpose into routine because it requires time and human resources.

There is less research on PTSD amongst children. However, researchers have collected several types of clinical and biological data amongst

children aged 7–18 years who were hospitalized following severe injuries [17]. The ability to predict PTSD at 3 months based on early data was accurate regardless of the method used and similar to previous results ($AUC = 79\%$ when combining all variables and $AUC = 74\%$ after feature selection).

PTSD is not often studied in relation to a specific trauma, but rather to “trauma in general.” This can lead to bias when interpreting and comparing studies. Few authors focus on a well-defined trauma and its association with PTSD. One study, however, focused on the trauma produced by exposure to combat in a context of armed conflict. Researchers were able to establish good performance in predicting PTSD on clinical and psychological data collected prior to military deployment ($AUC = 84\%$), while the integration of data collected after deployment only marginally increased the performance of the prediction model ($AUC = 88\%$) [18].

In addition to clinical and biological data that can be easily collected in hospitals or outpatient clinics, some authors have examined the use of neuroimaging data and smartphone questionnaire data for the early prediction of PTSD in different contexts: Li et al. (2016) obtained accurate performances in predicting PTSD at 6 months using neuroimaging data collected between 10 and 20 days after the trauma (Accuracy = 76%, Se = 73%, Sp = 76% with 49% PTSD prevalence after 6 months) [19]. And Wshah et al. (2019) collected mainly psychological data the first month after the trauma using a brief questionnaire sent to a smartphone, by combining an abbreviated 8-item version of PTSD Checklist for DSM-5 (PCL-5) and an additional item from PCL-5 assessing sleep with a tenth item assessing pain, managing to predict PTSD within 1 month, with the best performance for a scale using only 7 items ($AUC = 89\%$) [20].

Lastly, studies conducted outside hospitals are limited to mainly retrospective cohorts and secondary analyses. To our knowledge, no studies have focused on the prediction of PTSD in non-hospital contexts.

Data from retrospective studies based on the WHO WMH Surveys has been used to predict

PTSD, with one study focusing on a large number of traumas and another on trauma caused by sexual violence. They are based on a limited number of determinants (pretraumatic and peritraumatic factors, with trauma characteristics for the sexual violence study) [21, 22]. Based on these initial data alone, the results were variable, with excellent predictive performance in the study on all types of trauma ($AUC = 98\%$ in Kessler 2014) and poor for the sexual violence study ($AUC = 64\%$ in Scott 2018). The prediction algorithms used were not the same in both studies, with the superlearner set of algorithms giving better predictive performance than the single algorithm (logistic regression) used by Scott et al. Kessler et al. constructed a prediction model for any type of trauma but the challenge is to transpose these results to prospective data for external validation on just one type of trauma, since the prediction model for all types of trauma from a large cohort is not easily transposable to a smaller population or to a limited number of traumas. Similarly, two retrospective studies from large cohorts that each focused on a traumatic event (soldiers exposed to combat and earthquake victims) used initial data to predict PTSD (i.e., characteristics present before the trauma and those collected in the immediate aftermath in the case of the earthquake victims) [23, 24]. The predictive performance obtained was either good, in earthquake victims ($AUC = 79\%$ with a set of algorithms or super learner) or excellent, in the military ($AUC \geq 89\%$ with algorithms used separately).

Prediction of Response to Treatment

Once PTSD has been diagnosed and treatment is required, AI can also be employed to provide doctors with tools to help predict response to treatment. The use of AI is still rare in this field, with only two studies identified. The studies use neuroimaging data to predict the response to treatment in patients suffering from PTSD, but the very small number of subjects limits the relevance of the results. Zandvakili et al. used early (pre-treatment) EEG data to compare its capacity to predict response to treatment by PTSD patients and patients with depression who had been

exposed to combat [25], the treatment applied was repeated transcranial magnetic stimulation targeting identified neural network abnormalities: a treatment validated to treat pharmaco-resistant depression, still in the experimental stage for PTSD. Results indicated poorer performance in PTSD patients ($AUC = 71\%$) than in depressed patients ($AUC = 83\%$). Another example, Yuan et al. used functional MRI data prior to treatment to predict the response to an antidepressant treatment, Paroxetine, in earthquake-affected PTSD patients [26]. The predictive capacity of the Paroxetine efficacy at 3 months was accurate with $AUC = 72\%$ ($Acc = 72.5\%$, $Se = 67\%$, $Sp = 77\%$).

AI, Characterization and Diagnosis of PTSD. AI for Basic Research

The main focus of AI use in PTSD is for diagnosis. This can take two specific forms. It can be used to:

- (i) Provide diagnosis as a complement or substitute for measurement criteria or scales, should AI have at least one of the two advantages of (a) greater performance (sensitivity, specificity, accuracy) or (b) greater practical use and wider accessibility (e.g., less costly in terms of time, resources, and training of professionals). Making a diagnosis may also involve distinguishing it from diagnoses of other disorders, as PTSD, particularly in its chronic form, is frequently associated with other mental pathologies;
- (ii) Characterize the different initial and progressive forms of PTSD over time, as well as their relationship and combination with other disorders, especially mental disorders. This would be not so much a question of individualizing and personalizing diagnoses as of trying to identify whether there are particular types and subtypes of PTSD in order to ensure they are not ignored and to provide differential care and treatment if the associated consequences (response to treatment, duration of the disorder, severity of the

functional repercussions, associated disorders, etc.) differ significantly.

The majority of studies that use AI for characterizing PTSD (differential diagnoses, subtypes) are based on the use of genomic data and, particularly, neuroimaging. This can be explained by the need to associate types with a neuroanatomical or neurofunctional substrate, but the approach may prove to be limited in the sense that it does not take into account the broader syndromic expression of possible different types of PTSD. Another current limitation would be in the usefulness of this type of approach in routine clinical practice and primary care: the use of genomics or neuroimaging is not part of standard clinical examinations — and unlikely to be so for some time to come. These studies should therefore essentially be seen more as basic research, in advance of possible clinical applications.

The Use of Genomic and Neuroimaging Data by AI

Genomic Data

A person's genetic make-up may go some way to explaining their vulnerability to PTSD. A number of genes involved in the hormonal regulation of stress response (hypothalamic-pituitary-adrenal axis) have been identified as associated with the disorder. They are not specific to PTSD, however, as they are also found in other psychiatric disorders. One study on combat-exposed soldiers has used another panel of genes involved in immune and inflammatory system regulation as a genomic substrate to predict PTSD within 3 months of the soldier's return from deployment [27]. Data was taken from a blood sample. Although only a small number of soldiers were included in the study, two distinct prediction models based on the level of genomic expression, gene-expression, and exon-expression, respectively, showed good predictive performance, particularly the second model (Acc = 90%, Se = 100%, Sp = 80%).

Neuro-Imaging Data

Exposure to a traumatic event, particularly if complicated by PTSD, will have an impact on the

structure and functioning of the brain, mobilizing two types of substrate: neuroanatomical and neurofunctional. Once already established, PTSD is more difficult to detect and evaluate. This can be explained by the absence of pathognomonic symptoms and the presence of symptoms common to other psychiatric disorders. The evolution of PTSD is marked by the frequent appearance of mental disorders, with individuals with PTSD being more likely than those without PTSD to have symptoms that meet the diagnostic criteria for at least one other mental disorder (e.g., depression, anxiety disorder, bipolar disorder, etc.). This overlap is confusing for the clinician who needs to attribute symptoms to one or another disorder in an objective manner. In this context, neuroimaging can help to clarify the diagnosis of PTSD, as the various neuroimaging techniques employed pick up different levels of identifiable substrates. Research on the use of AI in this context has been limited to a limited number of traumatic events (road accidents, natural disasters, and exposure to combat), so the observations are valid for these events and are not necessarily transposable to other traumatic events likely to lead to PTSD. A series of examples have been selected for analysis in greater detail below in order to illustrate how the contribution of AI in neuroimaging can be used to help characterize PTSD in a variety of situations.

AI for the Diagnosis and Differentiation of PTSD from Other Mental Illnesses

Below we look at two specific examples that use retrospective data to compare different groups of exposed individuals with and without PTSD and mental illness: (i) differentiation of PTSD from the syndromic expression of the consequences of head injury [28]; (ii) differentiation of PTSD from depression [29].

(i) Mild traumatic brain injury can occur in a variety of traumatic situations, such as when soldiers are exposed to combat. Genuine PTSD can develop simultaneously with mild brain injury and further treatment can be complicated by the presence of neurocognitive symptoms which may be attributed to either.

In this study, AI and neuroimaging based on one type of neurofunctional substrate was 84% accurate in distinguishing soldiers with PTSD from soldiers with PTSD and brain injury.

- (ii) Depression is often associated with PTSD, suggesting a common susceptibility to both disorders. Symptoms of depression may overlap with symptoms with depressive connotations defined in Criterion D of DSM-5 for PTSD, when the symptoms of the latter are predominant, so impeding a diagnosis of PTSD. In this specific context, one study used neuroimaging and neurofunctional substrates to identify individuals with PTSD (especially PTSD associated with depression), achieving good predictive ability (AUC = 85% and Accuracy = 77%).

In both of these cases, the studies on AI coupled with neuroimaging supports the idea that a diagnosis of PTSD may be made some time after the traumatic event. Neuroimaging makes it possible to distinguish a disorder from others with which it is associated due to its intrinsic evolution. This clarification of the boundaries benefits the patient, for whom a customized diagnosis can be proposed with a focus on the therapeutic management that will result.

The Use of AI to Characterize Subtypes or Subsyndromic Forms of PTSD

The dissociative subtype of PTSD was included for the first time in the current version of the DSM-5. It defines a person who persistently or repeatedly experiences one of two dissociative symptoms: depersonalization and derealization. The dissociative form does not constitute a diagnostic entity that is independent or separate to standard PTSD; it is still PTSD, and to which can be added the two dissociative symptoms already contained in the PTSD criteria, i.e., flashbacks and dissociative amnesia). The dissociative form is relatively frequent, as according to the WHO WMH Surveys, it has been found in almost 14% of individuals with PTSD [30]. The distinctiveness of the dissociative form has also been highlighted by neuroimaging studies which have shown that there

are specific neurological pathways in subjects with dissociative symptoms. This data suggests that the distinct group with the dissociative form of PTSD possesses its own neurobiological and epidemiological characteristics. Neither the PCL-5 nor the CAPS-5 interview – two types of PTSD evaluation scale – are specifically designed to detect the dissociative form of PTSD. Neuroimaging coupled with AI can be used in prediction to compensate for this lack. This has been shown in two studies carried out by the same researchers. The studies had similar characteristics, differing only in the choice of neurofunctional substrates (balanced accuracy = 80%, in Nicholson 2020; balanced accuracy = 85–92% in Nicholson 2019) [31, 32].

Similarly, neuroimaging can also help to identify relationships between neurofunctional correlates and clusters of PTSD symptoms, those defined in criteria B to E of the DSM-5. In addition to the standard method of establishing a diagnosis by satisfying the required criteria, AI and neuroimaging would make it possible to identify subsyndromic forms of PTSD and to identify the class of dominant symptoms on the basis of neurofunctional correlates. Regression algorithms coupled with neuroimaging have shown that there are associations between neurofunctional correlates and PTSD symptom clusters, including avoidance ($R^2 = 0.23, p = 0.034$) and reexperiencing symptoms ($R^2 = 0.29, p = 0.002$) [33].

The Contribution of AI in Linking Basic Research to Clinical Applications. Neuroimaging Data

Most studies that have used AI coupled with neuroimaging (with magnetic resonance imaging) for diagnosing PTSD rely on group comparisons between healthy individuals and/or individuals who have suffered trauma but have not been diagnosed with PTSD. The studies have similar characteristics in terms of the type of data collection (retrospective), the type of traumatic event studied (limited to combat, earthquakes and road accidents), and the small size of the study group (57–140 individuals) [34–37]. The types and combinations of neuroimaging data input into prediction models varied according to the studies,

with a combination of the two types of correlates (neuroanatomical and neurofunctional) for one and use of one type of correlate for the others (neuroanatomical or neurofunctional). Predictive performance was good to excellent when predicting PTSD among healthy individuals (e.g.: accuracy = 91% in Gong 2014; AUC = 90% and accuracy = 89% in Zhang 2016), compared to the population exposed to trauma, where they were either, at best, accurate (e.g.: accuracy = 67% in Gong 2014; AUC = 72% and accuracy = 68% in Zhang 2016; accuracy = 77% in Salminen 2019) or inconclusive in one case (accuracy = 55% in Li 2020).

In order to diagnose PTSD, other researchers have combined AI with another type of functional neuroimaging technique, magnetoencephalography, which measures the electrical activity of neurons whose synchronous neural interactions constitute a different type of PTSD neurofunctional substrate [38, 39]. Predictive performance for identifying PTSD among combat-exposed male soldiers was excellent in these studies (AUC = 90% in Zhang 2020; AUC ≥ 93% in James 2015).

A final example of the possible use of AI based on functional neuroimaging data is for predicting the evolution of PTSD over time [40]. The authors of one five-year study followed 30 subjects who had survived a fire in the underground, all of whom developed PTSD at the first assessment at 1–2 months. Models for predicting the evolution of PTSD in the medium (about 1.5 years after the trauma) and long term (about 2.5 years after the trauma) discriminated little when their performance was related to the prevalence of PTSD in these intervals, with accurate performance for fairly high prevalence (AUC = 73% medium term with PTSD prevalence of 77%; AUC = 77% long term with PTSD prevalence of 48%). The last very long-term prediction model was insignificant for PTSD prevalence of 12%.

Potential Trends and Future Challenges

There are still few studies on the contribution of AI to PTSD, but they are mainly focused on two fields: prediction in a clinical context, in particular

the early prediction of PTSD, and prediction of response to treatment. AI has been used in basic research on genomic data, and more particularly, on anatomical or functional neuroimaging, either to identify specific substrates of PTSD in relation to other mental pathologies, or to explore the finer detail of the structure of PTSD, in its initial clinical variety or in its different evolutions over time.

There seems, therefore, to be a desire to use AI to help redefine or refine the diagnosis of PTSD, in particular by basing it on anatomical, physiological, and clinico-biological (including genetic) findings and correlates. No really significant finding has yet been made. The search for typologies, variants and syndromic trajectories in the same disorder is typical in mental health research, but it rarely involves the majority of the research community. And while the identification of subtypes may be of interest in itself, it is often limited in practical terms, as the subtypes are not necessarily easy to identify in an already difficult main diagnosis which is often overlooked, or they are not necessarily associated with significantly different consequences or treatments, if they even exist at all. Lastly, the majority of studies conducted to date focus on the use of AI in replacing current tools that are difficult for practitioners to use (complementary examinations that are difficult to access or interpret, and evaluation scales) or tools whose initial design, validation and use are the result of research (scales, DSM criteria). Nevertheless, these studies are mostly limited to traumas of all types, for people in emergency departments or hospitals, and do not allow us to be categorical about their real usefulness. They do, however, offer reasonable hope for more affordable, perhaps more effective and more operational means for health professionals to be able to diagnose PTSD.

The studies available employ mainly secondary use data, i.e., data from epidemiological cohorts originally constituted for other purposes, or reconstituted hospital cohorts. This data is clinical and biological, sometimes genomic, and very occasionally “real life” data collected via smartphones. The use of neuroimaging data is more specific to the field of mental health and is reserved for upstream research. The use of electronic medical records in outpatient clinics has not been explored to our knowledge.

Most of the studies, including those concerned with the use of AI in a hospital context, are related to laying the groundwork and providing prototypes that will enable the future conception of algorithms for general use. We are therefore only at the very beginning of the process of rigorous development of predictive or decision-support algorithms in this field. All the standard questions are still open with regard to the following steps. These include issues to do with the real possibility of integrating the algorithm in practitioners' working environments and information systems; resistance to variability in the quality of the data collected, among systems and among observers; prospective independent validation of algorithms on other data; the real acceptability and appropriation of this algorithm by the health professionals concerned. More fundamentally, the possible biases found in standard studies on PTSD and the collection of data used in the development of algorithms must be systematically detected and compensated for. This is the case for the most common biases encountered: those linked to ethnicity, geographical origin, gender, and age. Furthermore, the effort undertaken in other sectors of biomedical research to improve the quality and usefulness of research, for example, through randomized trials (CONSORT) or observational studies (STROBE) and through sufficient reporting to ensure the transparency of research and its reproducibility, should be replicated in research on AI. With regard to ongoing research and its goals (predictions and search for typologies), it would be mainly a matter of aiming for essential improvements: in experimental design (cross validation, prospective validation, transition to real life, and clinical tests, etc.); in reporting (metrics selected, a minimum of respect for TRIPOD guidelines, etc.); in research on the different traumatic events rather than a mixture; and in research on the prevalence of the outcome in the samples used).

AI seems promising as it provides the opportunity to unravel and describe existing interrelationships in terms of multiple causalities and consequences, both amongst symptoms constituting PTSD well as amongst associated pathologies (depression, for example), regardless of the nature of their relationship (simple association, common

mechanisms leading to two pathologies, consequences of each other, one variant of the other, etc.).

In all cases and whatever the scenario, there should be research and support on the place of AI and its acceptance by health professionals, while the use of AI could be extended to those who are not trained doctors, such as trained nurses, or even administrative staff in some cases, and possibly those working on traffic accidents, the military, etc. AI could be used as an alternative to health professionals when there is a shortage of qualified personnel or if training costs are high; it could also be used by those who are less qualified than specialists, for diagnosis, screening, or referral in the care system; and it could also help specialists themselves. Regardless of the professional using these algorithms, there will be an imperative need for high quality data collection (from medical observations and symptom collection) to input into the algorithm.

References

1. Liu H, Petukhova MV, Sampson NA, et al. Association of DSM-IV posttraumatic stress disorder with traumatic experience type and history in the World Health Organization World Mental Health Surveys. *JAMA Psychiatr.* 2017;74:270–81.
2. Galatzer-Levy IR, Bryant RA. 636,120 ways to have posttraumatic stress disorder. *Perspect Psychol Sci.* 2013;8:651–62.
3. Tolin DF, Foa EB. Sex differences in trauma and post-traumatic stress disorder: a quantitative review of 25 years of research. *Psychol Bull.* 2006;132:959–92.
4. Brewin CR, Andrews B, Valentine JD. Meta-analysis of risk factors for PTSD in trauma exposed adults. *J Consult Clin Psychol.* 2000;68:748–66.
5. Ozer EJ, Best SR, Lipsey TL, Weiss DS. Predictors of posttraumatic stress disorder and symptoms in adults: a meta-analysis. *Psychol Bull.* 2003;129:52–73.
6. Tortella-Feliu M, Fullana MA, Pérez-Vigil A, et al. Risk factors for posttraumatic stress disorder: an umbrella reviews of systematic reviews and meta-analyses. *Neurosci Biobehav Rev.* 2019;107:154–65.
7. Pitman RK, Rasmusson AM, Koenen KC, et al. Biological studies of posttraumatic stress disorder. *Nat Rev Neurosci.* 2012;13:769–87.
8. Rosellini AJ, Liu H, Petukhova MV, et al. Recovery from DSM-IV post-traumatic stress disorder in the WHO World Mental Health surveys. *Psychol Med.* 2018;48:437–50.
9. Kessler RC, Sonnega A, Bromet E, et al. Posttraumatic stress disorder in the National Comorbidity Survey. *Arch Gen Psychiatry.* 1995;52:1048–60.

10. Wilcox HC, Storr CL, Breslau N. Posttraumatic stress disorder and suicide attempts in a community sample of urban American young adults. *Arch Gen Psychiatry*. 2009;66:305–11.
11. Pacella ML, Hruska B, Delahanty DL. The physical health consequences of PTSD and PTSD symptoms: a meta-analytic review. *J Anxiety Disord*. 2013;27:33–46.
12. Schultebraucks K, Shalev AY, Michopoulos V, et al. A validated predictive algorithm of post-traumatic stress course following emergency department admission after a traumatic stressor. *Nat Med*. 2020;26:1084–8.
13. Papini S, Pisner D, Shumake J, et al. Ensemble machine learning prediction of posttraumatic stress disorder screening status after emergency room hospitalization. *J Anxiety Disord*. 2018;60:35–42.
14. Galatzer-Levy IR, Ma S, Statnikov A, et al. Utilization of machine learning for prediction of post-traumatic stress: a re-examination of cortisol in the prediction and pathways to non-remitting PTSD. *Transl Psychiatry*. 2017;7:e0.
15. Karstoft KI, Galatzer-Levy IR, Statnikov A, et al. Bridging a translational gap: using machine learning to improve the prediction of PTSD. *BMC Psychiatry*. 2015;15:30.
16. Galatzer-Levy IR, Karstoft KI, Statnikov A, Shalev AY. Quantitative forecasting of PTSD from early trauma responses: a machine learning application. *J Psychiatr Res*. 2014;59:68–76.
17. Saxe GN, Ma S, Ren J, Aliferis C. Machine learning methods to predict child posttraumatic stress: a proof of concept study. *BMC Psychiatry*. 2017;17:223.
18. Karstoft KI, Statnikov A, Galatzer-Levy IR. Early identification of posttraumatic stress following military deployment: application of machine learning methods to a prospective study of Danish soldiers. *J Affect Disord*. 2015;15:170–5.
19. Li L, Sun G, Liu K, et al. White matter changes in posttraumatic stress disorder following mild traumatic brain injury: a prospective longitudinal diffusion tensor imaging study. *Chin Med J*. 2016;129:1091–9.
20. Wshah S, Skalka C, Price M. Predicting posttraumatic stress disorder risk: a machine learning approach. *JMIR Ment Health*. 2019;6:e13946.
21. Kessler RC, Rose S, Koenen KC, et al. How well can posttraumatic stress disorder be predicted from pre-trauma risk factors? An exploratory study in the WHO World Mental Health Surveys. *World Psychiatry*. 2014;13:265–74.
22. Scott KM, Koenen KC, King A, et al. Post-traumatic stress disorder associated with sexual assault among women in the WHO World Mental Health Surveys. *Psychol Med*. 2018;48:155–67.
23. Rosellini AJ, Dussaillant F, Zubizarreta JR, et al. Predicting posttraumatic stress disorder following a natural disaster. *J Psychiatr Res*. 2018;96:15–22.
24. Leighton D, Williamson V, Darby J, Fear NT. Identifying probable post-traumatic stress disorder: applying supervised machine learning to data from a UK military cohort. *J Ment Health*. 2018;28:34–41.
25. Zandvakili A, Philip NS, Jones SR, et al. Use of machine learning in predicting clinical response to transcranial magnetic stimulation in comorbid post-traumatic stress disorder and major depression: a resting state electroencephalography study. *J Affect Disord*. 2019;252:47–54.
26. Yuan M, Qiu C, Meng Y, et al. Pre-treatment resting-state functional MR imaging predicts the long-term clinical outcome after short-term paroxetine treatment in post-traumatic stress disorder. *Front Psych*. 2018;9:532.
27. Tylee DS, Chandler SD, Nievergelt CM, et al. Blood-based gene-expression biomarkers of post-traumatic stress disorder among deployed marines: a pilot study. *Psychoneuroendocrinology*. 2015;51:472–94.
28. Rangaprakash D, Deshpande G, Daniel TA, et al. Compromised hippocampus-striatum pathway as a potential imaging biomarker of mild-traumatic brain injury and posttraumatic stress disorder. *Hum Brain Mapp*. 2017;38:2843–64.
29. Zilcha-Mano S, Zhu X, Suarez-Jimenez B, et al. Diagnostic and predictive neuroimaging biomarkers for posttraumatic stress disorder. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2020;5:688–96.
30. Stein DJ, Koenen KC, Friedman MJ, et al. Dissociation in posttraumatic stress disorder: evidence from the world mental health surveys. *Biol Psychiatry*. 2013;73:302–12.
31. Nicholson AA, Harricharan S, Densmore M, et al. Classifying heterogeneous presentations of PTSD via the default mode, central executive, and salience networks with machine learning. *Neuroimage Clin*. 2020;27:102262.
32. Nicholson AA, Densmore M, McKinnon MC, et al. Machine learning multivariate pattern analysis predicts classification of posttraumatic stress disorder and its dissociative subtype: a multimodal neuroimaging approach. *Psychol Med*. 2019;49:2049–59.
33. Zandvakili A, Barredo J, Swearingen HR, et al. Mapping PTSD symptoms to brain networks: a machine learning study. *Transl Psychiatry*. 2020;10:195.
34. Li Y, Zhu H, Ren Z, et al. Exploring memory function in earthquake trauma survivors with resting-state fMRI and machine learning. *BMC Psychiatry*. 2020;20:43.
35. Salminen LE, Morey RA, Riedel BC, et al. Adaptive identification of cortical and subcortical imaging markers of early life stress and posttraumatic stress disorder. *J Neuroimaging*. 2019;29:335–43.
36. Zhang Q, Wu Q, Zhu H, et al. Multimodal MRI-based classification of trauma survivors with and without post-traumatic stress disorder. *Front Neurosci*. 2016;10:292.
37. Gong Q, Li L, Tognin S, et al. Using structural neuroanatomy to identify trauma survivors with and without post-traumatic stress disorder at the individual level. *Psychol Med*. 2014;44:195–203.

38. Zhang J, Richardson JD, Dunkley BT. Classifying post-traumatic stress disorder using the magnetoencephalographic connectome and machine learning. *Sci Rep.* 2020;10:5937.
39. James LM, Belitskaya-Lévy I, Lu Y, et al. Development and application of a diagnostic algorithm for post-traumatic stress disorder. *Psychiatry Res.* 2015;231:1–7.
40. Im JJ, Kim B, Hwang J, et al. Diagnostic potential of multimodal neuroimaging in posttraumatic stress disorder. *PLoS One.* 2017;12:e0177847.
41. American Psychiatric Committee on Nomenclature and Statistics. *Diagnostic and statistical manual of mental disorders.* 5th ed. Washington, DC: American Psychiatric Association; 2013.



D. Kopyto, L. Uhlenberg, R. Zhang, V. Stonawski, S. Horndasch,
and Oliver Amft

Contents

Introduction	1644
AI for ED	1644
Monitoring of Dietary Behavior	1645
Methods in Automated Dietary Monitoring (ADM)	1645
Wearable-Based ADM	1645
Smartphone-Based ADM	1648
Ambient Technology-Based ADM	1648
Digital Biomarkers for AIM in EDs	1649
Intake Timing	1649
Food Type	1650
Food Amount	1652
Intake-Accompanying Phenomena Related to EDs	1653
Triggers and Stressors	1653
Example: AI in Anorexia Nervosa (AN)	1653
References	1657

Abstract

Over the past decades, the burden of eating disorders (ED) and comorbidities increased worldwide. Assisting diet monitoring with AI

methods and Automated Dietary Monitoring (ADM) can support ED risk prediction, diagnosis, tracking associated symptoms, and medical guidance during a long-term behavior change process. This chapter gives an overview of important directions in AI in the field of EDs and obesity. State-of-the-art methods and technologies for ADM are summarized in connection to digital biomarkers that reflect diet-related behavior in general. Two sensor-based ADM examples are detailed: food type classification and eating timing estimation. On the example of anorexia nervosa (AN), diet-related psychological parameters are detailed

D. Kopyto · L. Uhlenberg · R. Zhang · O. Amft (✉)
Digital Health, Friedrich-Alexander University Erlangen-Nuremberg, Erlangen, Germany
e-mail: david.kopyto@fau.de; lena.uhlenberg@fau.de;
rui.rui.zhang@fau.de; oliver.amft@fau.de

V. Stonawski · S. Horndasch
UK Erlangen, Friedrich-Alexander University Erlangen-Nuremberg, Erlangen, Germany
e-mail: valeska.stonawski@uk-erlangen.de;
stefanie.horndasch@uk-erlangen.de

and AI-based approaches to support AN diagnosis and treatment are described.

Keywords

Wearable computing · Ubiquitous computing · Dietary behavior · Embedded systems · ADM · Ambient technology · Smart eyeglasses · Digital biomarkers

Abbreviations

ADM	Automated Dietary Monitoring
AI	Artificial Intelligence
AN	Anorexia Nervosa
AUC	Area Under Curve
BCT	Behavior Change Technique
BMI	Body Mass Index
BN	Bulimia Nervosa
BOW	Bag-of-Words
BPNN	Back Propagation Neural Network
CLEF	Conference and Labs of the Evaluation Forum
CNN	Convolutional Neural Network
CONSORT	Consolidated Standards of Reporting Trials
CV	Cross Validation
DoG	Difference of Gaussians
ECG	Electrocardiography
ED	Eating Disorders
EMA	Ecological Momentary Assessment
EMG	Electromyography
ERDE	Early Detection Error
FFNN	Feed Forward Neural Network
GBT	Gradient Boosted Trees
HMM	Hidden Markov Model
HOG	Histogram of Gradients
HR	heart rate
HRV	heart rate variability
IMU	Inertial Measurement Unit
LASSO	Least Absolute Shrinkage and Selection Operator
LOC	loss-of-control eating
LR	Logistic Regression
mEMDA	mobile device-assisted Ecological Momentary Diet Assessment
ML	Machine Learning
MRI	Magnetic Resonance Imaging
NLP	Natural Language Processing

NPV	negative predictive value
ocSVM	one-class Support Vector Machine
PCA	Principal Component Analysis
PLS	Partial Least Squares
PPV	positive predictive value
RBF	Radial Basis Function
RNN	Recurrent Neural Network
SIFT	Scale-Invariant Feature Transform
SVM	Support Vector Machine
UMLS	Unified Medical System

Introduction

To facilitate intervention and therapy in eating disorders (ED), long-term monitoring of the patient and their eating behavior is crucial. Traditionally, dietary monitoring has been done by diaries that the patient completed on a daily basis. This chapter addresses Automated Dietary Monitoring (ADM), which intends to replace manual data logging by sensor-based inference, as far as reasonable. ADM is motivated by the insight that diaries are challenging to maintain by respondents and suffer from over- and underreporting. However, accurate information on the individual's actual diet is key for dieticians, physicians, and psychologists to support behavior and lifestyle changes over months and years. While EDs, including anorexia nervosa (AN), have been traditionally considered separate from obesity, there are substantial commonalities regarding the relevance of psychosocial, biological, and genetic components [1]. In particular for disordered eating that is related to forms of undernutrition, identification of symptoms and diagnosis can be challenging. Consequently, another stream of AI-related research addresses the multilevel diagnostics and symptom interpretation in EDs. The chapter covers AI methods and Machine Learning (ML) approaches on EDs and malnutrition.

AI for ED

According to psychiatric classification schemes like the ICD-10 or the DSM-V [2, 3], EDs can be classified as AN, bulimia nervosa, and binge

eating. Disordered eating, especially bulimia nervosa and binge eating disorders, may be considered not only an important risk but also a maintaining factor for obesity [4, 5]. Hence, obesity can be a consequence of EDs, metabolic overweight, and unhealthy food. The increasing burden of EDs, obesity, and comorbidities worldwide [4] requires new inventions to tackle the problem. AI can have an impact on the field by improving effectiveness of medical diagnoses, interventions, and therapies. For example, ADM-based digital biomarkers could help patients to understand and track eating behavior, and potentially adapt lifestyle more easily. Similarly, dietitians could leverage ADM to provide more fine-grain, real-time guidance in daily life, compared to using questionnaires or diaries. Finally, AI can support diagnoses of EDs based on genetic and other data inputs.

Monitoring of Dietary Behavior

Wearables, including smartwatches, in-ear headphones (which come with microphones and other sensors), or smart eyeglasses (e.g. [6–8]), can be used to acquire different behavioral and physiological data, which is used to determine biomarkers. Figure 1 shows often used devices and sensors to monitor dietary behavior.

Following the processing pipeline (cf. Fig. 1), AI and ML methods include data segmentation into meaningful subsequences, feature extraction, inference, and the estimation of digital biomarkers, which in turn can be used as guidance information or to direct interventions. Starting with the segmentation, acoustic data acquired from bone-conducting sensors at the head could be divided into subsequences of individual chewing sequences, i.e. instances of bites being chewed. Segmentation can be a multi-stage process, e.g. the audio chewing sequences could be further segmented into individual chewing cycles, i.e. data subsequences extending from one mandible closing to the next. Automating segmentation can be referred to as the most challenging part of the ADM pipeline (e.g. Paessler et al. [9, 10], Amft et al. [11, 12]). Depending on the time series to be analyzed, features include maxima of the sequence, MFCC features, and edge-based features in images. The resulting feature vectors are used for

inference, and the result can be interpreted as digital biomarkers. The inference can be achieved by classifiers, including Support Vector Machines (SVMs), unsupervised techniques, e.g. clustering, and regression. The devices and sensors utilized in the ADM pipeline illustrated in Fig. 1 are discussed in section “[Methods in Automated Dietary Monitoring \(ADM\)](#),” and the process to derive digital biomarkers, such as chewing timing or nutrients, including processing steps from AI and ML is detailed in section “[Digital Biomarkers for AIM in EDs](#).”

Methods in Automated Dietary Monitoring (ADM)

ADM aims at retrieving digital biomarkers, including the consumed food type, nutritional information, and food amount with as few user interactions as possible. ADM helps to integrate dietary information to daily life routines, which gives to medical staff and coaches a more representative picture of the patient’s eating behavior. Toward a more complete impression of dietary routines, various technological approaches have been taken. Many dietary measurement approaches produce similar biomarkers but use different sensor types and recording devices, which imply distinct AI techniques for processing the data. The different ADM approaches are summarized, and the challenges associated with each technique are highlighted, considering (1) wearable-based, (2) smartphone-based, and (3) ambient-based ADM. For each of the three approaches, the most important surveys are presented, and differences between the methods are discussed.

Wearable-Based ADM

Wearable devices can measure time series that capture phenomena during eating.

Figure 2 provides an overview on typical wearable devices used for ADM. To retrieve digital biomarkers, like food type, eating timing, and intake, accompanying phenomena, different signal types, and wearables can be exploited. Amft [13] summarized on-body measurements scenarios, particularly, the result of the device positions and

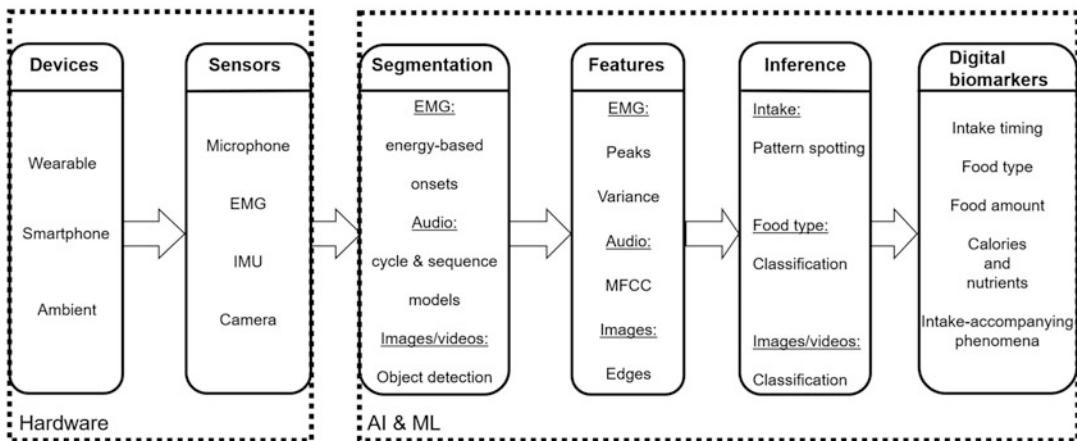


Fig. 1 The ADM data-processing pipeline with examples. Devices are equipped with sensors that record time series. The data is processed to yield feature vectors, which are used for inference of digital biomarkers

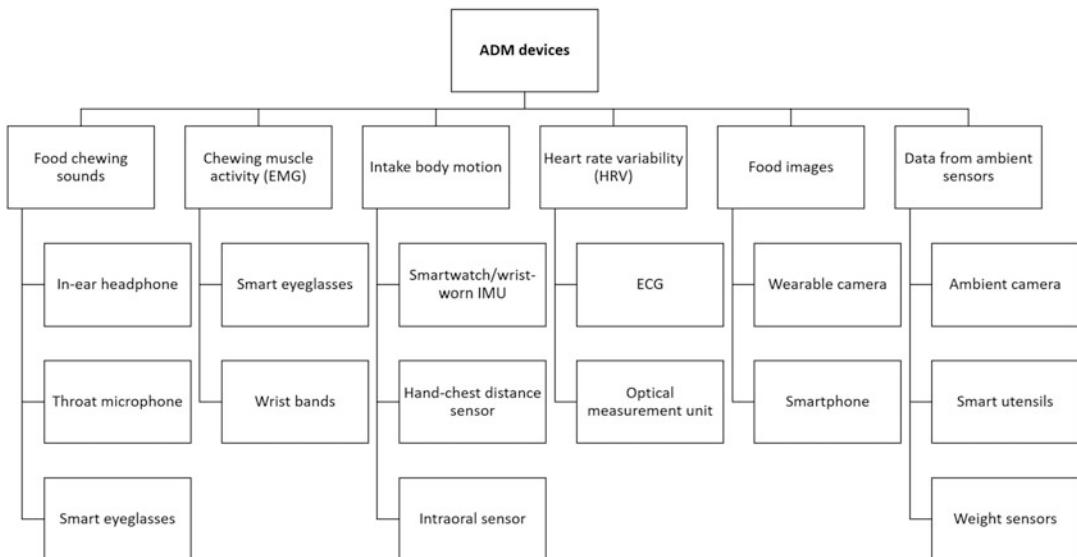


Fig. 2 An overview of frequently considered devices for ADM. Devices may be equipped with several sensor types

phenomena that can be acquired at the body positions. Eating Monitoring was grouped into different stages. Intake motions, i.e., the process of maneuvering a food item into the mouth, can be captured by inertial sensors or magnetic position sensors at wrists and chest. Inertial sensors can be used for chewing assessment too, by measuring head movement amplitudes. A variety of wearable sensors were found to capture chewing activity, including capacitive sensors, electromyographic (EMG), acoustic sensors, and dental implants.

Deglutition can be assessed with similar sensor types. The survey by Amft also addresses the use of temperature sensors to measure thermogenesis in response to food intake, and electro-gastrographic measurements to capture peristalsis movement to propel and process food. Heart rate and blood pressure sensors can be used to monitor diet-related physiology. Body composition is captured by bioimpedance sensors in supine position.

Schiboni and Amft [14] provided a comprehensive overview of the different wearable-based

ADM techniques, and expert models that describe food processing. The authors compared ear-worn devices to eyeglasses, mouth-embedded sensors and throat microphones in the context of chewing monitoring. Eyeglass-based dietary assessment includes EMG monitoring (Zhang et al. [6, 15–17]), piezoelectric strain (Farooq et al. [18]), load cells (Chung et al. [19]), and vibration sensors (Zhang and Amft [20]). Smart eyeglasses can be used for eating timing detection, as well as food type classification. Schiboni and Amft point out that full integration of these sensors into one wearable device is a benefit of this approach, as they are invisible to the user, while wearing during the day is comparatively convenient. The review paper also covers ear-worn devices. Ear-worn devices, such as in-ear headphones with integrated microphones are mostly used to record audio data [10, 12, 21–23]. The audio data coming from ear-worn acoustic sensors can be used for many different digital biomarkers, including eating timing, food type, intake quantity, or hydration timing. Audio data enables analysis of chewing microstructure. Since ear-worn acoustic sensors are located in the same place as hearing aids or earphones, it is convenient for patients to use them. On the downside, in-ear devices can occlude hearing, and their acoustic signals are perturbed by ambient noise. Intraoral sensors include inertial sensors [24] or strain gauge [25] and are mainly used for eating timing detection. Intraoral sensors can directly capture chewing motions. However, intraoral sensors are invasive, e.g. artificial teeth with embedded sensors require surgery, temporary implants may change oral sensation and behavior, and there is a risk that the sensor might be swallowed by the user. Another audio-based approach are throat microphones. Bi et al. [26] as well as Olubanjo and Ghovanloo [27] used this technology for eating timing detection. The main advantages of throat microphones, according to Schiboni and Amft [14], are the accessibility of both mastication and deglutition sounds, and the location can be also used for other physiological measurements. But microphones are also perturbed by ambient noise, speaking, and neck movement. The throat microphones require tight sensor-skin contact, and their

position is not ideal for wearable accessories and convenient use in daily living. Schiboni and Amft provide a similar comparison of approaches for ADM techniques dedicated to deglutition analysis. Furthermore, the survey provides insight about the use of wearable cameras to monitor food consumption of patients. In visual eating scene analysis, a camera is mounted to the user chest, head via cap, etc. to capture egocentric video and images. The resulting data can be used to get a more complete picture of intake events and nutritional data of the user. It has to be noted that continuous recording of camera footage causes privacy concerns due to unwanted images included in the material. Schiboni et al. investigated egocentric camera position further [28]. The authors implemented a camera on a cap's visor pointing downward. As a result, the unwanted privacy-threatening content decreased. Furthermore, the researchers proposed a dietary event-spotting algorithm based on video data coming from the device.

Vu et al. [29] presented another review of wearable techniques in ADM. Acoustic, visual, inertial, EMG/ECG, and piezoelectric sensing is covered. Apart from discussing advantages and disadvantages as well as current applicability and challenges, the authors outline the user comfort of different devices. Inner-ear devices for acoustic measurements are, for instance, classified as moderately comfortable, whereas visual approaches are seen as highly comfortable for the user. Information fusion approaches are an important focus of this chapter. Information fusion is considered as combining two or more wearable approaches in ADM. Combining sensors can increase accuracy or facilitate signal processing. As a prominent example, the authors point out the combination of acoustic and visual sensing. Liu et al. [30] designed a device, which embeds both a camera and a microphone for dietary assessment. As another sensor fusion idea, the survey of Vu et al. [29] mentions the possibility of combining inertial and visual sensors. Sen et al. [31] used the inertial measurement unit (IMU) of a smartwatch combined with a smartphone's camera that is triggered by the IMU data.

Smart eyeglasses with integrated sensors is a promising approach of using wearables in ADM. Introduced by Zhang and Amft [8], the idea has been utilized by different researchers. Zhang and Amft incorporated EMG electrodes and a measured chewing from muscle activity [20] and a vibration sensor to measure the skull vibration during chewing. More details of their approach are also illustrated in section “[Digital Biomarkers for AIM in EDs](#).” Furthermore, Wahl et al. [32] placed an accelerometer and gyroscope, as well as an optical heart beat sensor (photoplethysmography), into the eyeglasses frame. The setup enables investigators to detect daily activities, i.e., brushing teeth, reading, cycling, and jogging, which could be incorporated to assess daily activity of, e.g. obesity patients to complement their nutrition diary.

Smartphone-Based ADM

Sen et al. [31] used smartphones for dietary assessment. The authors combined smartwatch-based monitoring with a smartphone that communicates with an image recognition server. Adding a smartphone to the ADM-processing chain has been tried by various research groups.

Schembre et al. [33] summarized app-based approaches for dietary monitoring. Their focus is on mobile device-assisted ecological momentary diet assessment (mEMDA). mEMDA is an extension of ecological momentary assessment (EMA), also known as “in the moment assessment” or “experience sampling” [34]. The survey by Schembre et al. summarizes mEMDA protocols that have been used in research. The authors conclude that mEMDA compared to EMA can reduce participant burden and recall bias. An overview of several smartphone-based studies is given. The authors incorporated 20 studies using unique mEMDA protocols. A majority of these protocols (60%) asked the study participants to report their dietary data on the smartphone. Others were Internet-based applications. Some apps use social media (Twitter) to retrieve data. Sample periods vary from 3 days to 3 months, where most are in the range of 1 week. Data collection is done in three different ways: Some applications use

image-assisted dietary records, where images are taken before and after the meal with fiducial markers. Others prefer dietary records, where food ought to be chosen from different options or food groups. One approach used voice-annotated videos including time stamps. Data processing and nutrient analysis were also addressed. Furthermore, the review mentions smartphone-based nutrient analysis.

Villinger et al. [35] looked at app-based eating analysis too. The survey aimed at assessing the usability of methods to improve nutrition behaviors and related indices, including body mass index (BMI) and blood lipid concentration. The authors searched seven databases from 2006 to 2017. They included 41 studies and assessed them in a heatmap regarding the 25 Consolidated Standards of Reporting Trials (CONSORT) criteria. The survey quantifies behavior change technique (BCT) across the different studies and shows a forest plot that illustrates the effects of app-based mobile interventions on nutrition behaviors and other nutrition-related health results. Furthermore, the authors quantify the effects of app-based intervention on short-term, intermediate-term, and long-term follow-up intervals.

Ambient Technology-Based ADM

Sensor devices in the dining room, objects, e.g. cutlery, a fridge, etc., have been considered to track dietary habits, which Amft [13] summarized as ambient ADM technology and techniques. The survey distinguishes different targets to be monitored by the ambient sensors. Routine monitoring by looking at recurrent daily activities can be achieved by RFID-tags or a sound sensor [36]. Tags might be placed at arbitrary objects, while the RFID reader must be worn by the patient. Food preparation may be guided by recommendations generated by the ambient ADM system. To design an ambient technology-based recommender system, RFID, weight sensors, cameras, force sensors, or magnetic sensors could be integrated in kitchen trays, utensils, and furniture [37]. Intake monitoring, which is one of

the main applications of wearable-based ADM, can also be implemented in an ambient context. RFID-tags, weight sensors, or cameras are the preferred sensors and may be integrated into a dining table. Weight sensors could also be embedded into plates to measure food weight. Behavior tracking using vision-based observation, namely by surveillance video installations, has been used, e.g. at ceilings [38, 39]. Another option is to automatically register food products, which could be done by smart cards or RFID-tags put onto the products in the refrigerator or kitchen cabinets with the reader close by [40].

Digital Biomarkers for AIM in EDs

Intake Timing

The life of ED patients is often accompanied with irregular meal times due to, e.g. skipping meals or overeating. Irregular meal times and snacking could be revealed by intake time. Intake time can be detected with several sensor modalities. Amft et al. [11] applied an in-ear microphone for ADM. An eating detection algorithm based on the C4.5 Decision Tree classifier yielded an accuracy of 99%. Farooq et al. [18] used a temple-attached piezoelectric strain sensor and an accelerometer integrated in eyeglasses. A two-stage eating detection algorithm (Support Vector Machine and Decision Tree) yielded an average F1 score of 99% and area under the curve (AUC) of 0.99. Dong et al. [41] applied a naive Bayes classifier on wrist-worn motion data (accelerometer and gyroscope), yielding 81% accuracy, 0.6 min average intake start error, and 1.5 min end error. Thomaz et al. [42] used wrist-worn accelerometers to capture intake gestures in free-living with a Random Forest classifier, reporting 66% precision and 88% recall. Bedri et al. [43] used a wearable device called EarBit integrated with IMU and proximity sensors to capture free-living data. Eating detection based on the Random Forest produced an accuracy of 95%. Zhang et al. [44] proposed a neckband integrated with IMU, ambient light, and proximity sensors for free-living eating monitoring. A F1 score of 77% was

achieved using an eating detection algorithm based on Friedman's Gradient Boosting Model. Most of the above works evaluated the retrieval performance yet rarely reported absolute intake timing errors.

Chewing is often the most time-consuming process during eating solid and semisolid food [45]. Therefore, chewing duration can be regarded as a proper approximation of intake time. An example of deriving intake timing using EMG sensors embedded in a smart eyeglasses frame is detailed below. Temporalis muscles contract and relax during chewing. Stainless-steel EMG electrodes were used bilaterally in temple ear bends of 3D-printed eyeglasses to capture the contraction of Temporalis muscles of the wearer, yielding two channels of EMG signals (see Fig. 3).

Various eating detection algorithms were investigated [6, 17]. Eating detection algorithms can be categorized into top-down (detect eating from data directly) and bottom-up (detect chewing then eating) strategies. For example, a top-down approach is to apply a sliding window on rectified EMG data and extract features for a one-class Support Vector Machines (ocSVM) classifier with the Radial Basis Function (RBF) kernel. A label (eating or non-eating) was predicted by the trained ocSVM classifier and the overlapping labels were merged by majority voting, resulting in chewing segments. Finally, the gaps between two adjacent chewing segments less than 5 min were filled since they were considered as the natural breaks during an ongoing meal. A gap-free eating segment was considered as an eating event with its start, end, and duration representing the intake time.

In the bottom-up detection approach, individual chewing cycles were first detected, and chewing segments derived by clustering temporally close by chewing cycles. Gap handling, identical as in the top-down algorithm, was applied to the merged eating segments, yielding predicted eating events.

Following the above example, algorithms were analyzed in a smart eyeglasses dataset from ten participants, containing 44 eating events (meals and snacks) with durations ranging from 47 s to 36.2 min, totaling 429 min. F1 score and timing

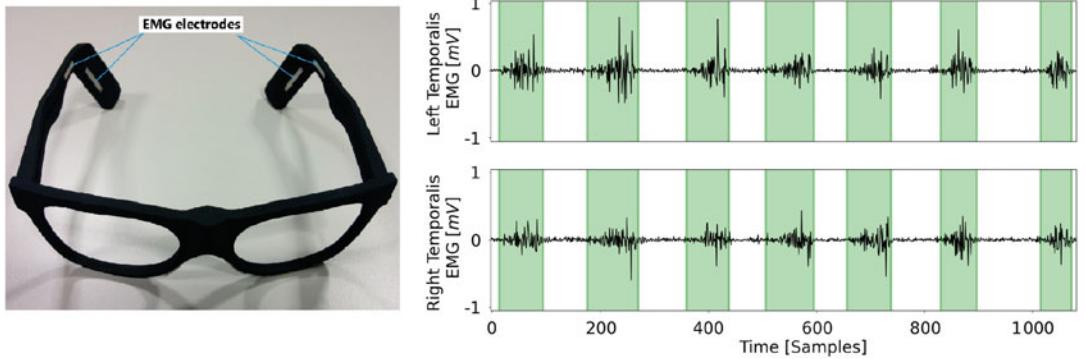


Fig. 3 Left: dietary monitoring smart eyeglasses integrated with bilateral EMG electrodes [16]. Right: example chewing EMG signal recorded with the eyeglasses. Labels indicate chewing cycles

errors were used to evaluate the chewing time detection algorithms. Precision and recall were computed as: precision = $\frac{T_{\text{tp}}}{T_{\text{ret}}}$ and recall = $\frac{T_{\text{tp}}}{T_{\text{gt}}}$, where T_{gt} was the summed duration of all eating events according to the reconstructed annotations, T_{ret} referred to the summed duration of all retrieved eating events, and T_{tp} was the sum of overlapping duration of retrieved eating events and reconstructed annotations.

Timing errors are computed as the difference between reconstructed annotations and the temporal boundaries of an eating event:

$$\Delta \bar{T}_s = \frac{\sum_{i=1}^M \min \left(\left| \hat{T}_s(j) - T_s(i) \right| \right)_{j=1,2,\dots,N}}{M}. \quad (1)$$

$$\Delta \bar{T}_E = \frac{\sum_{i=1}^M \min \left(\left| \hat{T}_E(j) - T_E(i) \right| \right)_{j=1,2,\dots,N}}{M} \quad (2)$$

where $\Delta \bar{T}_s$ is the start, and $\Delta \bar{T}_E$ the end timing error averaged across M eating events, and $\hat{T}_s(j)$ and $\hat{T}_E(j)$ are the start and the end of the retrieved eating event j , respectively. N is the number of retrieved eating events.

Zhang et al. [6, 17] reported performance on the above example data and detection algorithms. The top-down algorithm achieved $F1 = 95\%$ with $\Delta \bar{T}_s \approx 21.8\text{s}$ and $\Delta \bar{T}_E \approx 14.7\text{s}$. The bottom-up

algorithm achieved $F1 = 99\%$ with $\Delta \bar{T}_s \approx 2.4\text{s}$ and $\Delta \bar{T}_E \approx 4.3\text{s}$. In summary, chewing and intake time can be detected, e.g. from Temporalis EMG data, with timing errors of below 5 s in everyday life situations.

Food Type

Distinguishing food types, e.g. chips from apples or baguette, has been tackled by different ADM approaches and classification algorithms. In this section, an overview of food type recognition using image and audio data is given.

Image-Based Food Type Recognition

In image data, food is detected using segmentation and classification algorithms. Wen Lo et al. [46] surveyed different image processing pipelines. The authors show that image recognition techniques can be used for both, food type classification and food amount estimation. For both digital biomarkers, the authors distinguish between traditional, manually designed feature-based image processing using, e.g. Scale-Invariant Feature Transform (SIFT) features, classified by Support Vector Machines (SVMs), and deep learning approaches, e.g. Convolutional Neural Networks (CNNs). Among traditional models were nearest-neighbor classifiers and SVMs with different feature sets, including SIFT, Difference of Gaussians (DoG), or Histogram of Oriented Gradients (HOG). Deep Learning approaches mostly used

CNNs, including extensions, such as inception. Apart from self-recorded data, which several research groups used, e.g. Schiboni et al. [28], three different standardized datasets were used for image-based food type classification, including Food-101, UEC-FOOD100, and UEC-FOOD256.

Audio-Based Food Type Recognition

Different food types can also be inferred from audio data. Schiboni and Amft summarized audio-based food type recognition [14]. Breaking of the food objects in the mouth produces texture-dependent sounds that are characteristic for foods. Chips, for instance, are categorized as dry-crisp, whereas an apple has a wet-crisp sound texture (see Amft et al. [11]). Baguette, in contrast appears to be lower in amplitude and less “crisp.” Fig. 4 shows the difference between audio sequences of potato chips and baguettes.

The audio signal of food processing in the mouth contains chewing sequences from intake to deglutition with multiple chewing cycles. Chewing cycles are short periods starting by the

crushing of the food by the mandible until the next crushing-caused transient [14]. The difference between the audio signals of chewing cycles of potato chips and baguette is displayed in Fig. 5. Time series segmentation is one of the most challenging aspects of audio-based food type and amount estimation. For both, sequence and cycle level, effort has been spent to model the time series using patterns. Paessler et al. [9, 10] divided chewing sequences into 2–3 parts: begin, middle, and end phase. They designed a Hidden Markov Model (HMM)-based sequence model, which consisted of different HMMs, one for each phase. The transition between the phases was modeled by a Finite State Grammar. For chewing cycles, Amft et al. [11] proposed a model, where the chewing cycle audio data is divided into four phases:

1. Closing of the mandible, to crush food material
2. Pause
3. Opening of the mandible, where material is uncompressed
4. Pause

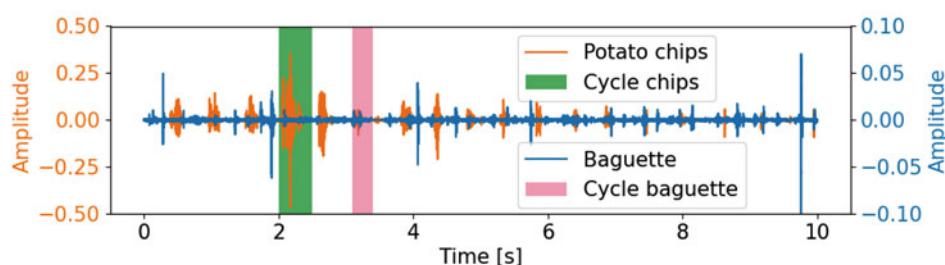


Fig. 4 Acoustic data sequences from two food samples. Potato chips have more prominent transients. Baguette transients are less prominent. One chewing cycle of potato chips and one chewing cycle of baguette are highlighted

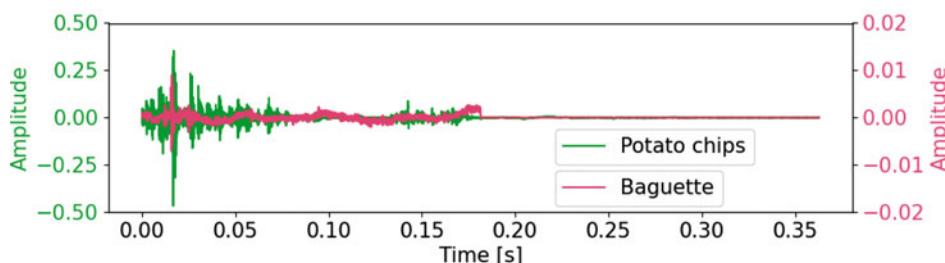


Fig. 5 Acoustic data from chewing cycles from two different food types. Potato chips have more prominent first and third phase. Baguette's amplitude is lower (max. displayed amplitude 0.02), and patterns are less distinct

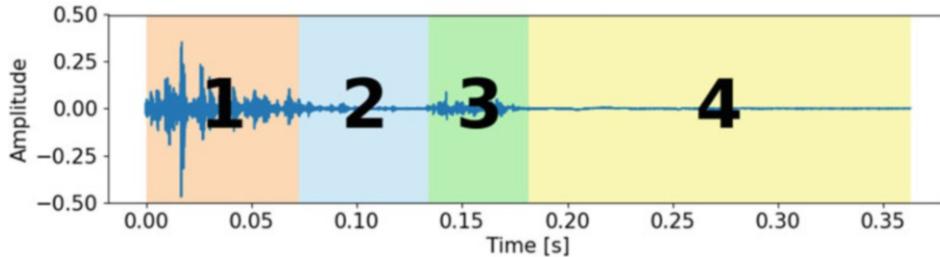


Fig. 6 Chewing cycle of potato chips with four phases

Figure 6 shows how the four phases are divided in a chewing cycle. Potato chips are a very illustrative example, where both the sequence model with begin, middle, and end phase as well as the four-phase cycle model can be applied. For other food types, where the amplitude of the transients is lower, the proposed sequence and cycle models are harder to infer (see the baguette cycle in Fig. 5). Segmentation was investigated by considering chewing cycles and swallowing events as symbols and adapting a probabilistic parser [47]. Another idea involves performing a multi-objective search to group chewing cycle patterns into subsequences [48]. Apart from time series segmentation, the audio data derived from acoustic sensors can be used for food type classification. Amft used a hierarchical approach [49] with a wearable earpad sensor to distinguish 19 food types. Furthermore, data of different sensor types were combined to determine food type. Mirtchouk et al. [22] used both in-ear audio data and wrist motion to classify food information. Lee et al. [50] used a Sonar placed close to the throat to record chewing signals. They extracted features and trained a neural network that could differentiate six food types. Mirtchouk et al. [22] distinguished 40 food types where audio and motion sensor data has been fused. The data was classified using hierarchical clustering.

Evaluation Metrics for Food Type Recognition

To evaluate algorithms that segment time series from food consumption or classify the different food types, different metrics need to be

considered. For example, Paessler et al. [10] reported classification by chewing sequence accuracy. For classification problems, the authors used accuracy to assess the quality of the outcome. Accuracy in this case is defined as:

$$\text{Accuracy} = \frac{|S_{\text{correct}}|}{|S_{\text{total}}|} \quad (3)$$

where S_{correct} is the set of all chewing sequences that have been correctly classified and S_{total} are the sequences in the dataset. A similar approach could be used for evaluating image classification algorithms.

Food Amount

ADM can determine the amount of consumed food in several ways. Wen Lo et al. [46] summarized how images can be segmented to estimate food amount. The different techniques include four categories: (1) Stereo-based approaches use multiple frames to reconstruct the 3D structure of food objects; (2) model-based techniques use pre-built mathematical models together with model scaling and rotation to determine the amount of food consumed; (3) others use depth-cameras, which can be used to determine the actual object scale without using fiducial markers, and perspective transformation refers to estimating object volume based on one single image; and (4) apart from the aforementioned techniques, deep learning models were used as well to estimate the volume of different foods.

Amft [13] looked at food amount estimation based on nonimage sensors. The estimation approach builds on the correlation of chewing cycles with food amount for a given food type and individual. Amft et al. [51] developed linear regression models to estimate food amount from chewing cycle count and food type information that was captured by an acoustic sensor. Prediction errors varied between 20% and 30% (Amft and Troester [23]). Food amount has also been estimated during drinking using magnetic position sensors [52]. Acoustic and EMG signal patterns during bolus swallowing were used to estimate fluid amount and viscosity too [53]. Further investigation by Mirtchouk et al. [22] confirmed and extended the food amount estimation principle by counting chewing cycles.

Determining calories and nutrients of consumed food is crucial to improve the lifestyle of patients. Schiboni and Amft [14] provided an overview of calories estimation methods. According to the authors, caloric amount can be inferred in different levels of granularity, e.g. calories per dietary event, and calories per bite. Dong et al. [54] used an indirect approach, where intake gestures were counted and a fixed calorie value per bite was considered. The authors' assumption is a simplification, as morsel weight or caloric value varies in foods and by the user's eating habits.

Calories are also retrievable from image data. Pouladzadeh et al. [55] applied using image processing and nutritional fact tables. Kirkpatrick and Collins [56] offered an introduction to nutrient estimation by ADM methods. Mankoff et al. [57] used shopping receipts to generate suggestions about healthier food items by scanning and passing them through an optical character recognition system. Rahman et al. [58] used the photo-acoustic effect to characterize liquid food in terms of their nutritional properties. They proposed a mobile system called Nutrilyzer. Anthimopoulos et al. [59] proposed a computer vision system to estimate carbohydrate consumption of diabetes type I patients using food images from a smartphone. Food images were segmented and

recognized using texture features and a nonlinear SVM. Carbohydrates were estimated by simulating the volume with a 3D model of the foods. Given the estimated volume, carbohydrates were determined from a dietary database.

All of the methods involving manual labor, e.g. picturing food, scanning receipts, etc., do not ensure that food amount was actually consumed by one individual.

Intake-Accompanying Phenomena Related to EDs

Triggers and Stressors

Excessive consumption of fatty foods and vitamin deficiencies due to an unbalanced diet are just two examples of malnutrition that can lead to serious health problems or be associated with EDs. Overall, there is little literature on AI regarding intake-accompanying phenomena. One of the most studied ED in this context is anorexia nervosa (AN), where AI methods have been applied. Possible triggers and stressors associated with EDs are manifold and can be related to, e.g. emotional problems, misperception of the body, or food stimuli (smells, e.g. of greasy food). Cravings in binge eating disorder, or avoidance of food in AN, may be triggered as a result. Furthermore, various authors have implicated that perceived stress may play an important role in the development of EDs (e.g. [60–62]). The following section "[Example: AI in Anorexia Nervosa \(AN\)](#)" will review AI-based approaches to support AN diagnosis and therapy. section "[Prospects for AN](#)" offers some insight on possible digital biomarkers for AN.

Example: AI in Anorexia Nervosa (AN)

AN is among the most common psychiatric ED. The prevalence accounts for 0.3% of all girls aged 13–18 years [63] and 1.1% of all female adults [64], with an increase in frequency over the past decade [65]. AN usually develops in early

adolescence and in 30–40% of patients becomes chronic leading to high morbidity and mortality [66]. However, diagnosing AN early is a major clinical challenge. Initial symptoms can be masked behind the pursuit of health or fitness goals, supported by family, peers, and health care professionals [67]. The field of possible application of AI in AN is broad, but not yet fully explored.

In 1998, Buscema et al. [68] made a first attempt to use ML methods to test the accuracy of neural networks in recognizing anorexic and bulimic patients. Data of 172 subjects with four different EDs was collected. The data base was composed of 124 variables, ranging from, e.g. generic information, alimentary behavior, menstrual cycles, weight, and height to psychodiagnostic tests. Six experiments were conducted, using a Feed Forward Neural Network (FFNN), with each of the six experiments using different variables. Decreasing the number of diagnostic conditions increased classification accuracy up to 100% for both AN and bulimia nervosa (BN) patients. Authors concluded that using variables from personal, social, and family data, chemical blood tests, and psychodiagnostic tests as input for the FFNN, it is possible to recognize AN and BN patients without knowing body weight and possible amenorrhea.

Guo et al. [69] also used ML methods in conjunction with genome data to predict the risk of AN. Whole genome genotyping data of 3940 AN cases and 9266 controls was analyzed. Model training was performed using a Lasso-regularized logistic regression (LR) model with tenfold Cross Validation (CV). For model testing and assessment, in addition to the LR model, Support Vector Machines (SVM) and Gradient Boosted Trees (GBT) were trained. A receiver operating characteristic curve (AUC) was derived to assess the performance of the different classifiers. The LR model generated an AUC of 0.693, while SVM and GBT reached AUCs of 0.691 and 0.623, respectively.

Furthermore, proof of the concept of an ML-based early-warning-system (ML-EWS) was

shown for AN patients [70], including 4,049 observations of 36 AN patients. Predictor variables of the ML-EWS models included 16 parameters ranging from, e.g. BMI, respiratory rate, oxygen saturation (SO₂%), heart rate, postural drop, blood investigations, and age. Predictive performance for Random Forests and naive Bayes classifier was assessed using 50-fold CV, yielding AUC of ~0.8 to 0.9. Despite performance drop from independent samples to multilevel analysis, multilevel RF slightly outperformed the existing EWS (AUC = 0.916), which did not use ML techniques.

Another interesting field for applying AI methods in AN is neuroanatomy [71, 72]. For other individual disease diagnoses, e.g. Alzheimer's and Parkinson's disease [72–75], ML techniques have been shown to reliably identify imaging biomarkers with an accuracy of above 90%. Lavagnino et al. [71] present a multivariate ML approach utilizing structural neuroanatomical scan data. T1-weighted magnetic resonance imaging (MRI) scans of 15 female AN patients and 15 demographically matched controls were acquired, and neuroanatomical volumes were extracted. The volumes served as input into the Least Absolute Shrinkage and Selection Operator (LASSO) multivariate ML algorithm with fivefold CV. Validity of the model was assessed using accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and AUC. Moreover, correlation of the results with “drive for thinness” and BMI were evaluated. The model achieved an accuracy of 83.3% (sensitivity 86.7%, specificity 80.0%). A linear relationship between drive for thinness clinical scores ($r = 0.52, p < 0.005$) and BMI ($r = -0.45, p = 0.01$) was shown.

Ceresa et al. [72] also assessed the feasibility of ML methods for extracting neuroimaging features. A total of 17 AN patients and 17 healthy controls were analyzed. Principal Component Analysis (PCA) was applied to structural T1-weighted MRIs. SVM algorithm was used to perform classification with 20-fold CV. Maps of the voxel-based pattern distribution of structural

brain differences were created. When using 31 PCA coefficients, accuracy, specificity, and sensitivity reached their best values of 0.85, 0.73, and 0.93, respectively. Consequently, using standard morphological brain images, SVM can extract neuroimaging biomarkers and accurately classify individuals with ED.

Zhao et al. [76] developed a predictive model for early diagnosis of AN based on the correlation between the concentration of six elements (Zn, Fe, Ca, Mg, Mn, and Cu), gender, and age. Data of 90 hair analyses (62 nonanorexic cases and 28 anorexic cases) was analyzed, and Partial Least Squares (PLS), Back Propagation Neural Network (BPNN), and SVM were applied. Classification accuracy revealed superior performance of SVM with an accuracy of 52%, 65%, and 87%, for PLS, BPNN, and SVM, respectively.

Another approach to predict risk or diagnose AN can be found in ML language-processing methods. Identification relies on medical reports, posted comments on the Internet, or self-reports of AN patients [77–79]. In 2018, Paul et al. [77] presented the results of using ada boost, LR, SVM, and RF classifiers on the Conference and Labs of the Evaluation Forum (CLEF) eRisk 2018 datasets. The authors used characteristic features obtained via Bag-of-Words (BOW) and Unified Medical Language System (UMLS) for model training and tenfold CV. The experimental analysis on the training set of AN patients showed that SVM classifier using BOW outperforms the other methods (precision = 0.97, recall = 0.98, F1 = 0.98, and ERDE₅₀ = 8.63%). In the same challenge, Wang et al. [78] use a classifier (TF-IDF) based on a Convolutional Neural Network (CNN) and weighting words with inverse document frequency. On the eRisk 2018 dataset, the model achieved decent Early Detection Error (ERDE₅) of 13.65%, ERDE₅₀ of 11.14%, and F1 = 0.67. Furthermore, Spinczyk et al. [79] used Natural Language Processing (NLP) methods in 44 female AN patients and presented a computer-aided ML-therapeutic diagnosis method for sentiment analysis of AN patient's statement. The intensity of five basic emotions

and six areas of difficulties [79] were analyzed by means of Natural Language Processing, Recurrent Neural Network (RNN) with two layers of Gated Recurrent Units [80], and Adam optimization. Compared to the dictionary-based method, the proposed RNN shows to be more effective (72% and 51%, respectively) in analyzing patient's body sentiment.

In summary, risk prediction has been the most widely used analysis objective in terms of AI in the field of AN. From the field of AI techniques, SVM, NN, PCA, and NLP techniques were successfully applied in risk assessment, establishment of early warning systems, and diagnosis of AN and showed promising results. Variables ranged from biological, genomic, to neuroanatomical parameters as well as language features from posts on the web or during therapy from the patients themselves.

Prospects for AN

Adolescent girls with AN undergo high levels of stress [81, 82], which has been shown as a risk factor in the development of EDs (e.g. [60–62]). Widely examined autonomic indices for stress reactivity are heart rate (HR) and heart rate variability (HRV, RR intervals (RRI) in electrocardiogram) [83]. A pilot study in adolescents showed associations between these two parameters and loss-of-control (LOC)-eating in the natural environment [84]. Peyser et al. [67] reviewed the current state of the art concerning HRV as a biomarker for AN and concluded that most studies show consistently elevated beat-by-beat variance for AN compared to controls. Thus, HRV features may act as a discriminating biomarker independent of low body mass or bradycardia [67] and may be used as input feature to different AI methods in the future.

Figure 7 shows an exemplary recording of RRI and derived HR. Measurements were performed using a wearable device worn on the biceps during a body exposure intervention session. Over the course of the intervention, the patient was confronted with a previously taken photo of her body twice. However, and in contrast to literature,

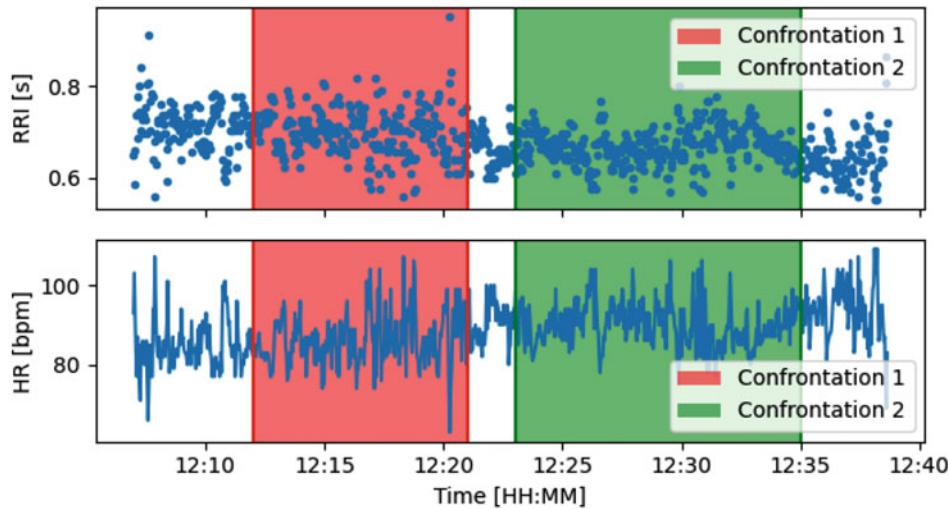


Fig. 7 Raw RRI and HR measurements during an intervention session. In contrast to literature, no clear HR increase or RRI decrease can be observed during body confrontation

no clear HR increase or RRI decrease was observed during the two body confrontation sessions, indicating that the variable relation may be inconsistent, even in controlled settings.

Furthermore, long-term measurements of HR and HRV in a controlled setting have not been conducted. Information on stress reactivity in general, and in relation to potentially stressful situations (e.g. social interactions or mealtimes) in particular, may be a future challenge for AI techniques in AN. Thus, possible correlations and risk predictions can be derived with different research objectives targeting AN.

Other promising biomarkers may be derived from activity monitoring as excessive physical activity is a common phenomenon in AN [85]. Enhanced activity could not only be used as a deliberate strategy to lose weight, but has also been discussed as an emotion regulation strategy [86]. Physical hyperactivity is an important prognostic factor in AN [87]. Correlations between physical hyperactivity and biological markers, including Leptin, Kisspeptin, and Ghrelin, have been shown [88]. However, there is no common definition of physical hyperactivity. Analyzing and classifying movement patterns and physical activity via means of AI methods has been

successfully shown in other clinical diseases, i.e. stroke and Parkinson's disease [89–91]. Applying AI techniques to AN motion data would offer the possibility of automatically detecting, extracting, and classifying movement patterns and other physiological features as biomarkers relevant to AN. Additionally, using these biomarkers as input to train adequate ML models could provide new tools to relate the biomarkers to ED-related psychopathology, to predict risk and disease/therapy progression, or to establish an early diagnosis.

As an example for motion data, Fig. 8 shows time series signals acquired by inertial measurement unit (IMU) sensors worn by the patient at each leg across a 12 h period. Labels for eating events and physical activity were taken from a paper protocol and superimposed on the signal. With the IMU data, AN-specific movement artifacts related to meals or leg bouncing, so-called feet fidgeting [92, 93], may be detected. AN-specific movements may potentially result from imposed stress and are just two examples of promising features that could be extracted and interpreted as possible digital biomarkers for AN. Measuring physical activity along with RRI and HR (see Fig. 8) may provide additional

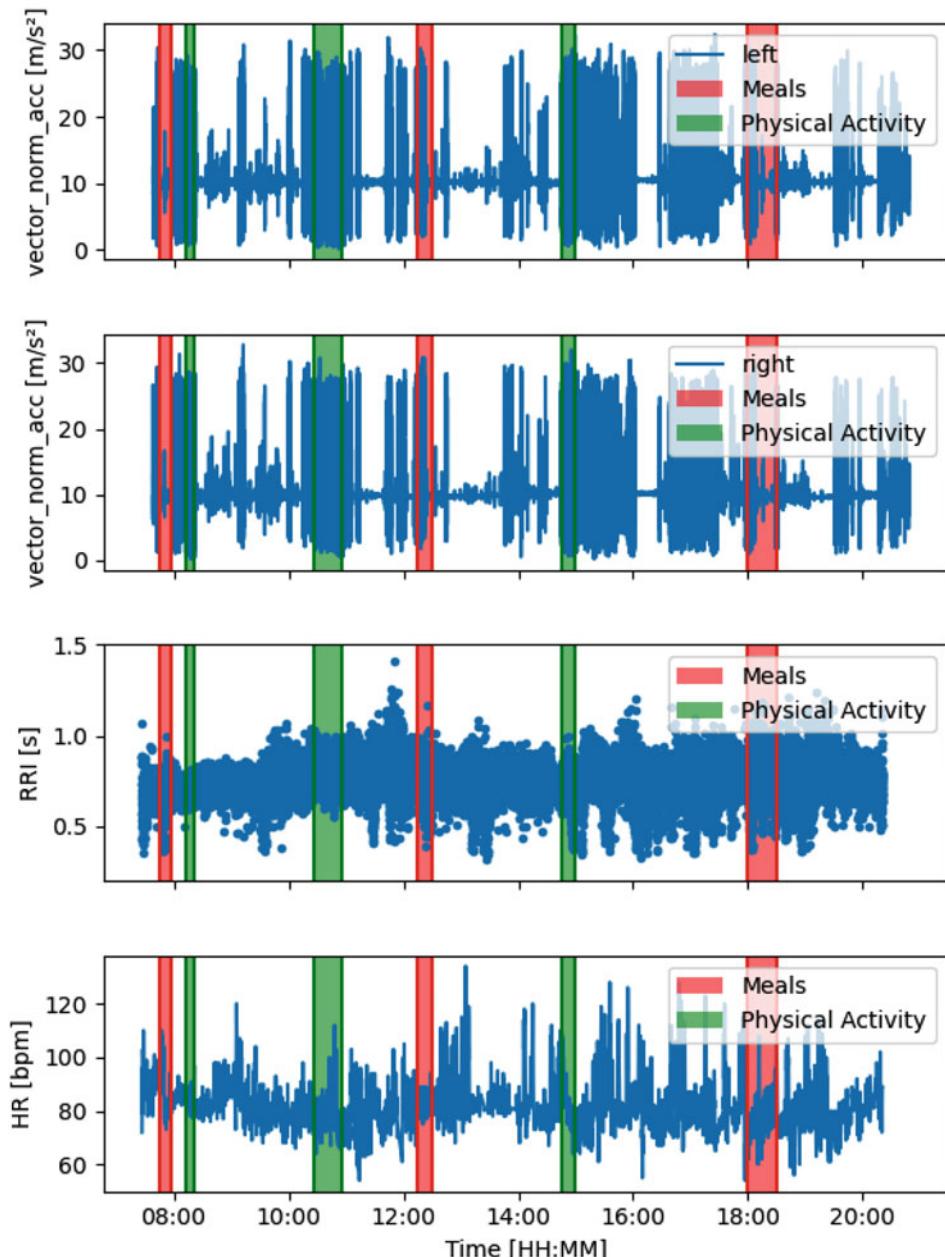


Fig. 8 Example of vector normalized time series signal of linear acceleration of both feet and cardiac measurements (RRI and HR) taken from the same subject and day

context information and enrich the input for different ML approaches.

Acknowledgements The work was partly supported by the BMBF Eghi project, #16SV8526.

References

- Day J, Ternouth A, Collier DA. Eating disorders and obesity: two sides of the same coin? *Epidemiol Psichiatr Soc*. 2009;18(2):96–100.

2. World Health Organization. International statistical classification of diseases and related health problems. 10th ed. Geneva: World Health Organization; 2015.
3. Diagnostic and Statistical Manual of Mental Disorders: DSM-5™, 5th Ed. Arlington, American Psychiatric Publishing, Inc.; 2013.
4. Hay P, Mitchison D. Eating disorders and obesity: the challenge for our times. *Nutrients*. 2019;11(5):1055.
5. da Luz FQ, Hay P, Touyz S, Sainsbury A. Obesity with comorbid eating disorders: associated health risks and treatment approaches. *Nutrients*. 2018;10(7):829.
6. Zhang R, Amft O. Retrieval and timing performance of chewing-based eating event detection in wearable sensors. *Sensors*. 2020;20(2):557.
7. Zhang R, Kolbin V, Süttenbach M, Hedges M, Amft O. Evaluation of 3D-printed conductive lines and EMG electrodes on smart eyeglasses frames. In: Proceedings of the 2018 ACM international symposium on wearable computers. ISWC '18. ACM, Singapore; 2018. p. 234–235.
8. Zhang R, Amft O. Regular-look eyeglasses can monitor chewing. In: Proceedings of the 2016 ACM international symposium on wearable computers (ISWC '16). ACM; 2016. p. 389–392.
9. Päßler S. Analyse des menschlichen Ernährungsverhaltens mit Hilfe von Kaugeräuschen /. vol. 72 of Studentexte zur Sprachkommunikation; 72. Dresden: TUDpress, Verl. der Wiss; 2014.
10. Päßler S, Wolff M, Fischer WJ. Food intake monitoring: an acoustical approach to automated food intake activity detection and classification of consumed food. *Physiol Meas*. 2012;33(6):1073.
11. Amft O, Stäger M, Lukowicz P, Tröster G. Analysis of chewing sounds for dietary monitoring. In: UbiComp 2005: Proceedings of the 7th international conference on ubiquitous computing. vol. 3660 of LNCS. Springer Berlin, Heidelberg; 2005. p. 56–72.
12. Amft O, Tröster G. Recognition of dietary activity events using on-body sensors. *Artif Intell Med*. 2008;42(2):121–36.
13. Amft O. Ambient, on-body, and implantable monitoring technologies to assess dietary behaviour. In: Preedy VR, Watson RR, Martin CR, editors. International handbook of behavior, food and nutrition, vol. 38. Springer, New York; 2011. p. 3507–26.
14. Schiboni G, Amft O. Automatic dietary monitoring using wearable accessories. In: Tamura T, Chen W, editors. Seamless healthcare monitoring: advancements in wearable, attachable, and invisible devices. Cham: Springer; 2018. p. 369–412.
15. Zhang R, Bernhart S, Amft O. Diet eyeglasses: recognising food chewing using EMG and smart eyeglasses. In: Proceedings of the international conference on wearable and implantable body sensor networks (BSN '16). IEEE; 2016. p. 7–12.
16. Zhang R, Amft O. Monitoring chewing and eating in free-living using smart eyeglasses. *IEEE J Biomed Health Inform*. 2018;22(1):23–32.
17. Zhang R, Amft O. Free-living eating event spotting using EMG-monitoring eyeglasses. In: Proceedings of the 2018 IEEE EMBS international conference on biomedical health informatics (BHI '18). Las Vegas: IEEE; 2018. p. 128–32.
18. Farooq M, Sazonov E. A novel wearable device for food intake and physical activity recognition. *Sensors*. 2016;16(7):1067.
19. Chung J, Chung J, Oh W, Yoo Y, Lee WG, Bang H. A glasses-type wearable device for monitoring the patterns of food intake and facial activity. *Sci Rep*. 2017;7:41690.
20. Zhang R, Amft O. Bite glasses: measuring chewing using EMG and bone vibration in smart eyeglasses. In: Proceedings of the 2016 ACM international symposium on wearable computers (ISWC '16). ISWC '16. New York: ACM; 2016. p. 50–2.
21. Gao Y, Zhang N, Wang H, Ding X, Ye X, Chen G, et al. iHear food: eating detection using commodity bluetooth headsets. In: Connected health: applications, systems and engineering technologies (CHASE), 2016 IEEE first international conference on IEEE; 2016. p. 163–172.
22. Mirtchouk M, Merck C, Kleinberg S. Automated estimation of food type and amount consumed from body-worn audio and motion sensors. In: Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing. UbiComp '16. New York: ACM; 2016. p. 451–62.
23. Amft O, Tröster G. On-body sensing solutions for automatic dietary monitoring. *Pervasive Comput*, IEEE. 2009;8(2):62–70.
24. Li CY, Chen YC, Chen WJ, Huang P, Chu H. Sensor-embedded teeth for oral activity recognition. In: Proceedings of the 2013 international symposium on wearable computers. ISWC '13. New York: ACM; 2013. p. 41–4.
25. Stellar E, Shrager EE. Chews and swallows and the microstructure of eating. *Am J Clin Nutr*. 1985;42 (5 Suppl):973–82.
26. Bi Y, Lv M, Song C, Xu W, Guan N, Yi W. AutoDietary: a wearable acoustic sensor system for food intake recognition in daily life. *IEEE Sensors J*. 2016;16(3):806–16.
27. Olubanjo T, Ghovanloo M. Real-time swallowing detection based on tracheal acoustics. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP); 2014. p. 4384–4388.
28. Schiboni G, Wasner F, Amft O. A privacy-preserving wearable camera setup for dietary event spotting in free-living. In: Proceedings of the international conference on pervasive computing and communications (PerCom) workshops; 2018. p. 872–877.
29. Vu T, Lin F, Alshurafa N, Xu W. Wearable food intake monitoring technologies: a comprehensive review. *Computers*. 2017;6(1):4.
30. Liu J, Johns E, Atallah L, Pettitt C, Lo B, Frost G, et al. An intelligent food-intake monitoring system using wearable sensors. In: 2012 ninth international

- conference on wearable and implantable body sensor networks (BSN); 2012. p. 154–160.
31. Sen S, Subbaraju V, Misra A, Balan RK, Lee Y. The case for smartwatch-based diet monitoring. In: 2015 IEEE international conference on pervasive computing and communication workshops (PerCom Workshops); 2015. p. 585–590.
 32. Wahl F, Zhang R, Freund M, Amft O. Personalizing 3D-printed smart eyeglasses to augment daily life. *IEEE Computer*. 2017;50(2):26–35.
 33. Schembre SM, Liao Y, O'Connor SG, Hingle MD, Shen SE, Hamoy KG, et al. Mobile ecological momentary diet assessment methods for behavioral research: systematic review. *JMIR Mhealth Uhealth*. 2018;6(11):e11170.
 34. van de Ven P, O'Brien H, Henriques R, Klein M, Msetfi R, Nelson J, et al. ULTEMAT: a Mobile framework for smart ecological momentary assessments and interventions. *Internet Interv*. 2017;9:74–81.
 35. Villinger K, Wahl DR, Boeing H, Schupp HT, Renner B. The effectiveness of app-based mobile interventions on nutrition behaviours and nutrition-related health outcomes: a systematic review and meta-analysis. *Obes Rev*. 2019;20(10):1465–84.
 36. Patterson DJ, Fox D, Kautz H, Philipose M. Fine-grained activity recognition by aggregating abstract object usage. In: Rhodes B, Mase K, editors. ISWC 2005: proceedings of the ninth IEEE International symposium on wearable computers. IEEE Press; 2005. p. 44–51.
 37. Chi PYP, Chen JH, Chu HH, Lo JL. Enabling calorie-aware cooking in a smart kitchen. In: Persuasive 2008: proceedings of the 3rd international conference on persuasive technology, vol. 5033. Oulu: Springer; 2008. p. 116–27.
 38. Hauptmann AG, Gao J, Yan R, Qi Y, Yang J, Wactlar HD. Automated analysis of nursing home observations. *IEEE Perv Comput*. 2004;3(2):15–21.
 39. Kakra V, van der Aa N, Noldus L, Amft O. A multimodal benchmark tool for automated eating behaviour recognition. In: Proceedings of measuring behavior 2014; 2014.
 40. Lambert N, Plumb J, Looise B, Johnson IT, Harvey I, Wheeler C, et al. Using smart card technology to monitor the eating habits of children in a school cafeteria: 1. Developing and validating the methodology. *J Hum Nutr Diet*. 2005;18(4):243–54.
 41. Dong Y, Scisco J, Wilson M, Muth E, Hoover A. Detecting periods of eating during free-living by tracking wrist motion. *IEEE J Biomed Health Inform*. 2014;18(4):1253–60.
 42. Thomaz E, Essa I, Abowd GD. A practical approach for recognizing eating moments with wrist-mounted inertial sensing. In: Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing. UbiComp '15. New York: ACM; 2015. p. 1029–40.
 43. Bedri A, Li R, Haynes M, Kosaraju RP, Grover I, Prioleau T, et al. EarBit: using wearable sensors to detect eating episodes in unconstrained environments. *Proc ACM Interact Mob Wearable Ubiquitous Technol*. 2017;1(3):37:1–37:20.
 44. Zhang S, Zhao Y, Nguyen DT, Xu R, Sen S, Hester J, et al. NeckSense: a multi-sensor necklace for detecting eating activities in free-living conditions. *Proc ACM Interact Mobile Wearable Ubiquitous Technol*. 2020;4(2):72:1–72:26.
 45. Schiboni G, Amft O. Sparse natural gesture spotting in free living to monitor drinking with wrist-worn inertial sensors. In: Proceedings of the 2018 ACM international symposium on wearable computers. ISWC '18. New York: ACM; 2018. p. 140–7.
 46. Wen Lo FP, Sun Y, Qiu J, Lo B. Image-based food classification and volume estimation for dietary assessment: a review. *IEEE J Biomed Health Inform*. 2020;24(7):1926–39.
 47. Amft O, Kusserow M, Tröster G. Probabilistic parsing of dietary activity events. In: Leonhardt S, Falck T, Mähönen P, editors. BSN 2007: Proceedings of the international workshop on wearable and implantable body sensor networks. vol. 13. Springer; 2007. p. 242–247.
 48. Amft O, Kusserow M, Tröster G. Automatic identification of temporal sequences in chewing sounds. In: Hu T, Mandoiu I, Obradovic Z, editors. BIBM 2007: proceedings of the IEEE international conference on bioinformatics and biomedicine. San Jose: IEEE Press; 2007. p. 194–201.
 49. Amft O. A wearable earpad sensor for chewing monitoring. In: Sensors 2010: Proceedings of IEEE sensors conference. IEEE; 2010. p. 222–227.
 50. Lee KS. Food intake detection using ultrasonic Doppler sonar. *IEEE Sensors J*. 2017;17(18):6056–68.
 51. Amft O, Kusserow M, Tröster G. Bite weight prediction from acoustic recognition of chewing. *IEEE Trans Biomed Eng*. 2009;56(6):1663–72.
 52. Amft O, Bannach D, Pirkl G, Kreil M, Lukowicz P. Towards wearable sensing based assessment of fluid intake. In: PerHealth 2010: Proceedings of the First IEEE PerCom workshop on pervasive healthcare. IEEE; 2010. p. 298–303.
 53. Amft O, Tröster G. Methods for detection and classification of normal swallowing from muscle activation and sound. In: PHC 2006: Proceedings of the first international conference on pervasive computing technologies for healthcare. ICST; 2006. p. 1–10.
 54. Dong Y, Hoover A, Muth E. A device for detecting and counting bites of food taken by a person during eating. In: Proceedings of the 2009 IEEE international conference on bioinformatics and biomedicine. BIBM '09. IEEE Computer Society, Washington, DC; 2009. p. 265–268.
 55. Pouladzadeh P, Shirmohammadi S, Al-Maghribi R. Measuring calorie and nutrition from food image. *IEEE Trans Instrum Meas*. 2014;63(8):1947–56.
 56. Kirkpatrick SI, Collins CE. Assessment of nutrient intakes: introduction to the special issue. *Nutrients*. 2016;8(4):184.

57. Mankoff J, Hsieh G, Hung HC, Lee S, Nitao E. Using low-cost sensing to support nutritional awareness. In: Goos G, Hartmanis J, van Leeuwen J, editors. Ubicomp 2002: Proceedings of the 4th international conference on ubiquitous computing. vol. 2498 of Lecture notes in computer science. Springer Berlin, Heidelberg; 2002. p. 371–376.
58. Rahman T, Adams AT, Schein P, Jain A, Erickson D, Choudhury T. Nutrilyzer: a mobile system for characterizing liquid food with photoacoustic effect. In: Proceedings of the 14th ACM conference on embedded network sensor systems CD-ROM. SenSys '16. ACM, New York; 2016. p. 123–136.
59. Anthimopoulos M, Dehais J, Shevchik S, Ransford BH, Duke D, Diem P, et al. Computer vision-based carbohydrate estimation for type 1 patients with diabetes using smartphones. *J Diabetes Sci Technol.* 2015;9(3):507–15.
60. Bittinger JN, Smith JE. Mediating and moderating effects of stress perception and situation type on coping responses in women with disordered eating. *Eat Behav.* 2003;4(1):89–106.
61. Sassaroli S, Ruggiero GM. The role of stress in the association between low self-esteem, perfectionism, and worry, and eating disorders. *Int J Eat Disord.* 2005;37(2):135–41.
62. Tozzi F, Sullivan PF, Fear JL, McKenzie J, Bulik CM. Causes and recovery in anorexia nervosa: the patient's perspective. *Int J Eat Disord.* 2003;33(2):143–54.
63. Swanson SA, Crow SJ, Le Grange D, Swendsen J, Merikangas KR. Prevalence and correlates of eating disorders in adolescents. Results from the national comorbidity survey replication adolescent supplement. *Arch Gen Psychiatry.* 2011;68(7):714–23.
64. Jacobi D, Perrin AE, Grosman N, Doräl MF, Normand S, Oppert JM, et al. Physical activity-related energy expenditure with the RT3 and TriTrac accelerometers in overweight adults. *Obesity.* 2007;15(4):950–6.
65. Micali N, Hagberg KW, Petersen I, Treasure JL. The incidence of eating disorders in the UK in 2000–2009: findings from the general practice research database. *BMJ Open.* 2013;3(5):e002646.
66. Dobrescu SR, Dinkler L, Gillberg C, Råstam M, Gillberg C, Wentz E. Anorexia nervosa: 30-year outcome. *Br J Psychiatry.* 2020;216(2):97–104.
67. Peyser D, Scolnick B, Hildebrandt T, Taylor JA. Heart rate variability as a biomarker for anorexia nervosa: a review. *Eur Eat Disord Rev.* 2021;29:20–31.
68. Buscema M, Pietralata MM, Salvemini V, Intraligi M, Indrimi M. Application of artificial neural networks to eating disorders. *Subst Use Misuse.* 1998;33(3):765–91.
69. Guo H, Chen L, Chen G, Lv M. Smartphone-based activity recognition independent of device orientation and placement. *Int J Commun Syst.* 2016;29(16):2403–15.
70. Ioannidis K, Serfontein J, Deakin J, Bruneau M, Ciobanca A, Holt L, et al. Early warning systems in inpatient anorexia nervosa: a validation of the MARSIPAN-based modified early warning system. *Eur Eat Disord Rev.* 2020;28(5):551–8.
71. Lavagnino L, Amianto F, Mwangi B, D'Agata F, Spalatro A, Zunta-Soares GB, et al. Identifying neuro-anatomical signatures of anorexia nervosa: a multivariate machine learning approach. *Psychol Med.* 2015;45(13):2805–12.
72. Cerasa A, Castiglioni I, Salvatore C, Funaro A, Martino I, Alfano S, et al. Biomarkers of eating disorders using support vector machine analysis of structural neuroimaging data: preliminary results. *Behav Neurol.* 2015;2015:924814.
73. Dyrba M, Ewers M, Wegrzyn M, Kilimann I, Plant C, Oswald A, et al. Robust automated detection of microstructural white matter degeneration in Alzheimer's disease using machine learning classification of multi-center DTI data. *PLoS One.* 2013;8(5):e64925.
74. Klöppel S, Stonnington CM, Chu C, Draganski B, Sechill RI, Rohrer JD, et al. Automatic classification of MR scans in Alzheimer's disease. *Brain.* 2008;131(Pt 3):681–9.
75. Cherubini A, Morelli M, Nisticó R, Salsone M, Arabia G, Vasta R, et al. Magnetic resonance support vector machine discriminates between Parkinson disease and progressive supranuclear palsy. *Mov Disord.* 2014;29(2):266–9.
76. Zhao CY, Zhang RS, Liu HX, Xue CX, Zhao SG, Zhou XF, et al. Diagnosing anorexia based on partial least squares, back propagation neural network, and support vector machines. *J Chem Inf Comput Sci.* 2004;44(6):2040–6.
77. Paul S, Kalyani J, Basu T. Early detection of signs of anorexia and depression over social media using effective machine learning frameworks; In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France; 2018.
78. Wang YT, Huang HH, Chen H. A neural network approach to early risk detection of depression and anorexia on social media text. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France; 2018.
79. Spinczyk D, Bas M, Dzieciątko M, Maćkowski M, Rojewska K, Maćkowska S. Computer-aided therapeutic diagnosis for anorexia. *Biomed Eng Online.* 2020;19(1):53.
80. Cho K, Merriënboer BV, Bahdanau D, Bengio Y. On the properties of neural machine translation: encoder-decoder approaches. In: Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar; 2014.
81. Beukens M, Walker S, Esterhuyse K. The role of coping responses in the relationship between perceived stress and disordered eating in a cross-cultural sample of female university students. *Stress Health.* 2010;26(4):280–91.
82. Tavolacci MP, Ladner J, Grigioni S, Richard L, Villet H, Dechelotte P. Prevalence and association of

- perceived stress, substance use and behavioral addictions: a cross-sectional study among university students in France, 2009–2011. *BMC Public Health.* 2013;13(1):724.
83. Kim HG, Cheon EJ, Bai DS, Lee YH, Koo BH. Stress and heart rate variability: a meta-analysis and review of the literature. *Psychiatry Investig.* 2018;15(3):235–45.
84. Ranzenhofer LM, Engel SG, Crosby RD, Haigney M, Anderson M, McCaffery JM, et al. Realtime assessment of heart rate variability and loss of control eating in adolescent girls: a pilot study. *Int J Eat Disord.* 2016;49(2):197–201.
85. Dalle Grave R, Calugi S, Marchesini G. Compulsive exercise to control shape or weight in eating disorders: prevalence, associated features, and treatment outcome. *Compr Psychiatry.* 2008;49:346–52.
86. Carrera O, Adan RAH, Gutierrez E, Danner UN, Hoek HW, van Elburg AA, et al. Hyperactivity in anorexia nervosa: warming up not just burning-off calories. *PLoS One.* 2012;7(7):e41851.
87. El Ghoch M, Calugi S, Pellegrini M, Milanese C, Busacchi M, Battistini NC, et al. Measured physical activity in anorexia nervosa: features and treatment outcome. *Int J Eat Disord.* 2013;46(7):709–12.
88. Hofmann T, Elbelt U, Haas V, Ahnis A, Klapp BF, Rose M, et al. Plasma kisspeptin and ghrelin levels are independently correlated with physical activity in patients with anorexia nervosa. *Appetite.* 2017;108:141–50.
89. Maceira-Elvira P, Popa T, Schmid AC, Hummel FC. Wearable technology in stroke rehabilitation: towards improved diagnosis and treatment of upper-limb motor impairment. *J Neuroeng Rehabil.* 2019;16(1):142.
90. Kubota KJ, Chen JA, Little MA. Machine learning for large-scale wearable sensor data in Parkinson's disease: concepts, promises, pitfalls, and futures. *Mov Disord.* 2016;31(9):1314–26.
91. Kobsar D, Ferber R. Wearable sensor data to track subject-specific movement patterns related to clinical outcomes using a machine learning approach. *Sensors (Basel, Switzerland).* 2018;18(9):2828.
92. Esseiva J, Caon M, Mugellini E, Khaled OA, Aminian K. Feet fidgeting detection based on accelerometers using decision tree learning and gradient boosting. In: Rojas I, Ortuño F, editors. *Bioinformatics and biomedical engineering*, vol. 10814. Cham: Springer International Publishing; 2018. p. 75–84.
93. Belak L, Gianini L, Klein DA, Sazonov E, Keegan K, Neustadt E, et al. Measurement of fidgeting in patients with anorexia nervosa using a novel shoe-based monitor. *Eat Behav.* 2017;24:45–8.



Daisy Das and Lipi B. Mahanta

Contents

Introduction	1664
Childhood Medulloblastoma	1665
Methods and Materials for AI Application	1665
Significance of an Accurate Detection in Prognosis	1670
Challenges of a Clinical Observation	1670
Advantages of AI-Driven Solution	1670
Challenges of an AI-Driven Solution	1671
Scope for Industry Transformation	1671
Summary	1671
References	1671

Abstract

This chapter describes the advancement of neurological treatment through artificial intelligence (AI)-driven tools and applications. It

AI as the protagonist for the better prognosis of childhood medulloblastoma with efficient subtype diagnosis

Supplementary Information: The online version of this chapter (https://doi.org/10.1007/978-3-030-64573-1_189) contains supplementary material, which is available to authorized users.

D. Das · L. B. Mahanta (✉)
Institute of Advanced Study in Science and Technology,
Guwahati, India
e-mail: lbmahanta@iasst.gov.in

introduces the need for AI research in neurology and a description of the components of AI for an AI-specific application. The role of AI in neurology and the mechanism of a machine learning (ML) or deep learning (DL) model can be found as the main context of the description. For a deeper level of understanding, this chapter also focuses on a case study that provides a stepwise guideline for the development of an AI application for the classification of medulloblastoma tumors to impart practical knowledge of the methods and methodologies involved in the formation of such AI models. Challenges of a clinical and AI-based observation for patient treatment, advantages of an AI-driven model, its need and usage, and industry transition are some discussions at the

later part of the chapter. Overall, this chapter focuses on the amalgamation of clinical observation with AI research and how AI can play a supportive role in neurological assessments.

Keywords

Medulloblastoma · AI · Neurology · Machine learning · Deep learning · Computer vision · AI diagnostics

Introduction

Artificial intelligence is the technique of feeding human decision-making or understanding capabilities to machines to mimic natural intelligence in problem-solving. AI has already been a solution to many natural object detection or scene understanding problems [1–3]. Recent advances in AI are seen in the healthcare sector that play a major role in the detection and identification of healthcare problems [4–6]. A report by [7] provides an overview of ML methods by neurosurgeons across the globe and its impact. As stated, AI has witnessed a recent rise with the advent of machine learning (ML) [8–12] and deep learning (DL) solutions particularly in the field of neurosurgery. ML and DL methods are domains of AI that use computer programs to automatically learn and identify problem data and give a predictive outcome. Until the introduction of AI in the field, the neurologist was the principal investigator of the diagnosis and prediction of a neurological disease involved in the decision-making process of any operative process, tumor resection, or post-operative risk and prognosis [11]. However, the individual decision of a neurosurgeon cannot be generalized to every patient because each patient has their own variant of patient history to account for. The number of patient records and information has given rise to an exponential rise of big data problems, which requires computer-based analysis and solutions. The processing of these huge patient records to give a predictive decision is beyond the scope of a medical practitioner. With the invasion of AI tools and applications in the medical and healthcare domain, it is possible and

necessary to assist neurosurgeons in integrating this patient information into reliable decision-making models for improved stratification.

AI generally comprises three learning models: supervised, unsupervised, and reinforcement. The supervised models are based on a training set of data with predefined classes of output where we know the actual class of the training data, and we need to train the machine based on this data to correctly identify the classes. On the other hand, unsupervised learning does not have labeled training data. Here the training data is automatically classified based on similarity measure that the training data exhibit among the same classes. Reinforcement learning is based on model feedback, where the model automatically learns to train itself based on the feedback it sends to itself after prediction, e.g., self-driving car. Generally, for medical data, the learning is supervised with prior knowledge and collected medical data.

[13–16] studies the impact of machine learning algorithms on neurosurgery data. The introduction of AI can generate the risk factors associated with any prognosis decision [17]. AI and its tools have a wide variety of applications in neurological treatment. A study was made where AI is used to access the suicidal risk factor and prevent suicide [18]. AI-based models have been studied to predict dementia, autism, and Alzheimer [19–23]. It is used to study neuro traumatology [24] to predict mortality associated with accidents and identify brain hemorrhages and strokes [25, 26] in neuro-imaging. It is used in neuro habitation for implementing robotic arms [27]. AI can be used to sense and analyze the patient data and generate a single decision-based system to give a predictive outcome integrating all the factors associated with the patient data [28, 29] to give the most favorable outcome. ML and DL algorithms are used in radiological or pathological image understanding for neurological data by image feature study [30, 31]. AI methods are also used in the preoperative stage, where information from a set of radiological images are preprocessed for better understanding using AI image processing methods or image segmentation techniques to get a clear understanding of the clinical region of interest from these images [32–34]. AI is also used

in making a postoperative decision by classifying and identifying the grading of a neurological tumor from a medical slide after a postoperative biopsy [35–37]. We can also find AI in the molecular study that could be useful in determining the pathogens in a neurosurgery sample [38, 39]. ML for neurology has a high range of publications from recent years [40–42]. The advent of AI in neurosurgery has made it possible to perform robotic surgery to assist the neurosurgeons [43]. AI-based ML models have also been used in the survival prediction of patients based on the grading and differentiation of tumor cells [44]. AI is also used in pathological scene understanding of the biopsy tissue samples for tumor grading and identification of a neurosurgery sample [45]. Modern methods of AI and data mining have already given pathways in generating prediction models for neurological oncology treatment [46]. Central nervous system (CNS) tumors (Fig. 1) are highly malignant and the leading cause of death among children [47]. As reported, nearly 4000 cases per year are diagnosed with these types of tumors in the United States, which causes 25% of cancer-related deaths. The overall incidence of brain tumors is highest in children with less than 5 years of age accounting for 75% of the cases in less than 10 years of age. The diagnosis of such tumors is difficult to obtain from children due to the difficulty in communication and clinical observation among infants and the prolonged delay for the implicative results of the tumor. There exists a substantial difference in the molecular biology of adults and children giving rise to the distinctive study of pediatric brain tumor. Astrocytoma, glioblastoma, oligodendrogiomas, ependymoma, and medulloblastoma are a few CNS tumors to name for [47, 48]. As defined, these tumors contain diverse pathological entities and are complicated for interpretation by a general pathologist and often require the guidance of an experienced neuropathologist as the intervention of samples between a general pathologist and a neuropathologist is high. The need for diagnosing such tumors is challenging, and this highlights the need for collaboration of multiple clinical research institute and computer scientist. In further reading, the steps in the application of AI

models in diagnosing these CNS pediatric brain tumor is explained with a special reference to childhood medulloblastoma tumor.

Childhood Medulloblastoma

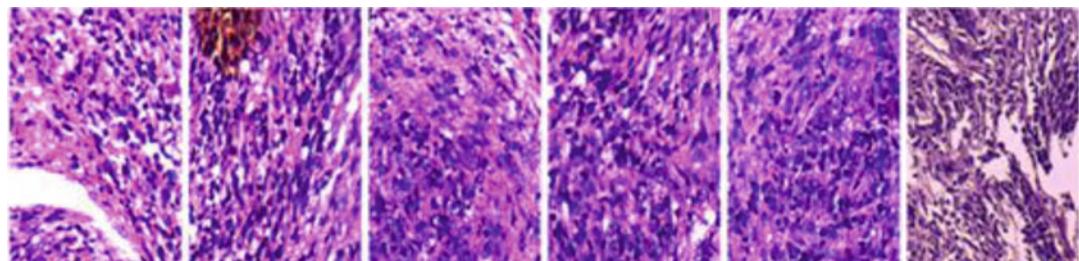
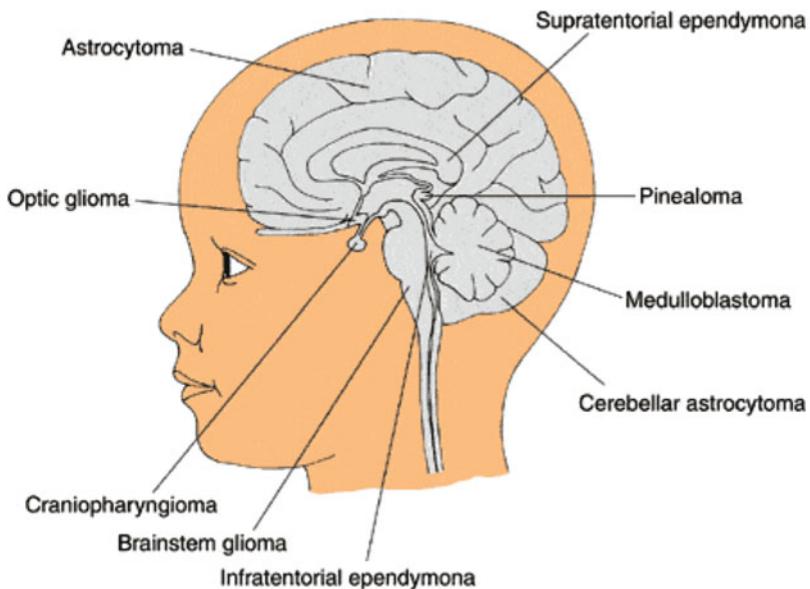
Childhood medulloblastoma (CMB) is a pediatric nervous system tumor that arises from the roof of the fourth ventricle in the cerebella hemisphere. Its peak incidence is seen in children from 5 to 6 years of age. The World Health Organization has graded it as a highly malignant grade IV tumor [49]. It has four different subtypes, namely, large cell, desmoplastic, classic, and nodular, each having a separate histological variant [50]. The author gives the pathological description of the different variants. The prognosis and survival rate depend on the different subtypes of the CMBs. The large subtype is highly anaplastic. The nodular subtypes have nodular patterns of cell arrangement, while the desmoplastic contains a network of collagen fiber, the classic subtype is most prevalent, and its cell arrangement can be microscopically found the same as the large except that it is non-anaplastic [51]. Macroscopic observations of these tumors are finely granular of grayish pink to purple with soft consistency accumulated lesion. Microscopically, these are carrot-shaped cells with scanty cytoplasm. Patients with CMB are often seen with headache, failure to thrive, cerebella symptoms, and nausea [52].

For malignant regions, it has a high density of cells with scanty cytoplasm and irregular cell structures, while in normal regions, the cells are not highly distributed and also have smooth round cells in shape.

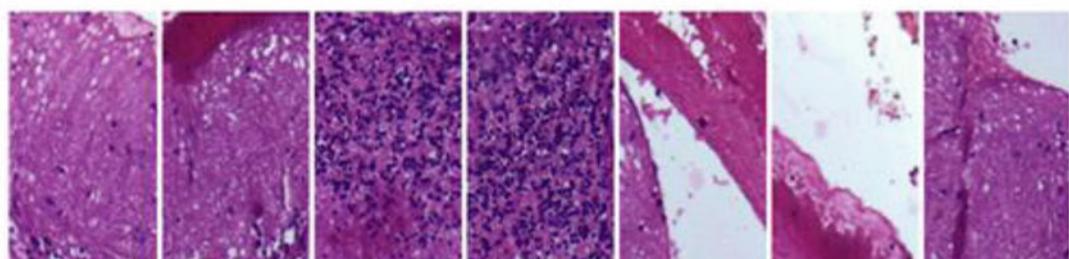
Methods and Materials for AI Application

Any ML-based classification for histopathological images will have the following steps, as shown in Fig. 4. To make it more clear, we discuss the following case study. The study [53–59] was made for biopsy images at both low power and high power microscopic observations, as

Fig. 1 The figure shows different types of CNS tumors with different region of origin. (Image courtesy: <https://childrenswi.org/medical-care/macc-fund-center/conditions/oncology/braintumors>)



a) Architectural view of malignant region



b) Architectural view of normal region

Fig. 2 Image showing the microscopic captured image of the (a) malignant regions and (b) normal cell regions in biopsy samples for brain tumors

observed by a pathologist for histopathological samples. The low power view would be referred to as an architectural level and high power as a cell level. The classification model was built for both

binary (Fig. 2) and multiclass (Fig. 3) at both levels. The binary classifier classified between normal and CMB sample [57], whereas the multiclass classified the subtypes of the CMB [56].

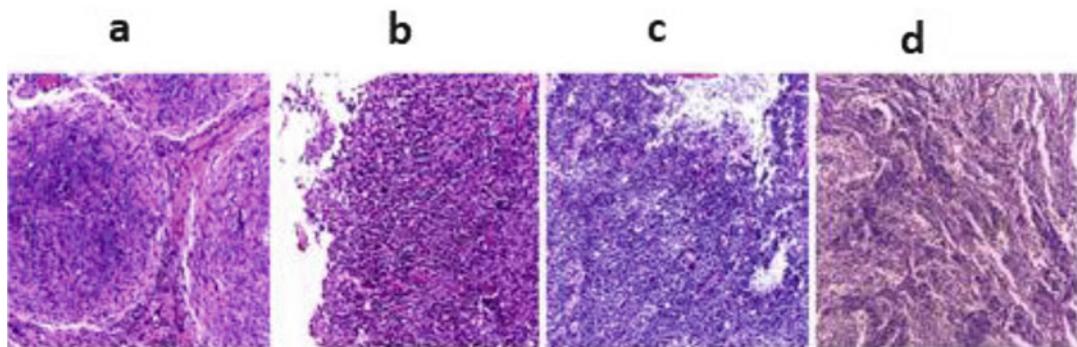


Fig. 3 Image showing the different subtypes of childhood medulloblastoma samples from microscopic image capture. (a) Nodular, (b) classic, (c) large cell, and (d) desmoplastic subtype

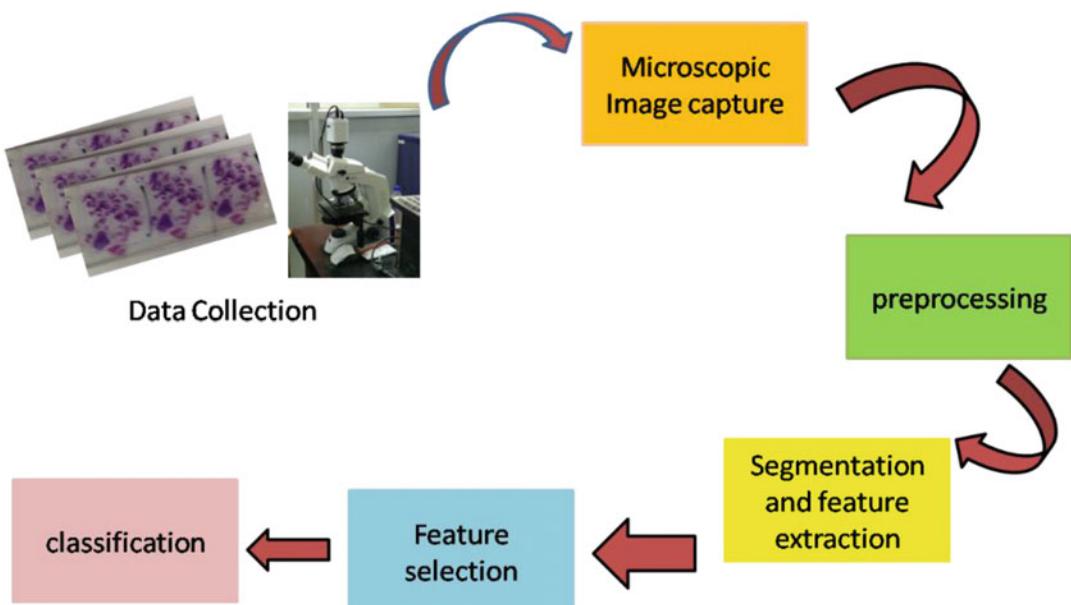


Fig. 4 Image classification steps for an ML classification

Data collection: The microscopic images are collected from the medical slides that were prepared from the tissue specimen collected from the postoperative procedure of tumor resection. The images were collected at $10\times$ and $100\times$ magnification in RGB standard color format and stored with a .jpg image extension [60].

Data preprocessing: Preprocessing is an additional step and is only done when the required information in the data needs to be highlighted or filtered for gaining insightful details from the concerned ROI. Preprocessing of the data was

done for $10\times$ images, which were used for classification at the architectural level. Color channeling was used for the identification of the color channel that gives us a good visualization of the tumor spread. The RGB image was decomposed into various color models and its constituent channels. Later, a color enhancement filtering on the selected color channel was performed.

Ground truth annotation: Any medical AI model would require the substantial support of ground truth for performance validation. This ground truth can only be marked and validated

by a clinical practitioner. However, the generation of ground truth data is a very tiring process and requires patience and devotion from a medical expert. A total of 1260 ground truth information was generated to identify the cells from the underlining smear images captured at $100\times$ [55]. The ground truth cell annotations were further processed by a machine learning K-means clustering algorithm to capture this ground truth and convert it into binary information for machine understanding. The ground truth marked cells are now identified by 1 (white) and the background as 0 (black). The ground truth information extraction is shown in Fig. 5.

Feature extraction: Texture study [56, 57] was made at the architectural level to study the structural arrangement of the tumor cells. Five different texture features were extracted. Both first-order and second-order features were studied. For first-order features, we considered a histogram-based feature of the pixel intensities that gave kurtosis, skewness, mean, probability of intensity value, etc. Total, there were 14 histogram-based features. For second-order intensity information, four features: i) the GLCM in four directions, ii) GRLM, iii) LBP, and iv) Tamura feature were used to understand the pixel relationship among the neighboring pixels. All total, it was 171 features for binary and 99 for multiclass. At the cell level, color and shape features were extracted in addition to these texture features. For a cell level feature understanding, the information from the ground truth annotated cells were used. The malignant cells tend to have a high color chromaticity and irregularity in shape as compared to normal cells, and so these were important

biomarkers for our study. All total, there were 259 cell features.

Feature selection: The selection of features is equally important as the extraction of important features. If the features are not valuable and do not contain a high set of information, this would adversely affect the classification performance. Stacking the classifier with unnecessary features will only increase the complexity of an ML model, decreasing the performance. For a DL network model, these features are automatically extracted by the model, while for ML methods, we have to perform a manual extraction and selection of features. A forward selection method was used for feature selection where at first, the individual accuracy of features was measured for classification, and then they were grouped in subgroups of features to extract the feature subset that carries maximum information.

Feature reduction: A high dimension of features has the probability of having a high amount of redundant features. In machine learning, there is an underlying curse of dimensionality associated with large feature sets that affect the performance of an ML classifier. For this, it is required to understand the contribution of each individual feature toward the identification of the classification problem. Feature reduction is the technique of reducing the dimension of a feature set without loss of information. Feature reduction was made using principal component analysis (PCA). The features obtained from feature selection were converted to their principal component and arranged in descending order of their eigenvalues. The accuracy of the classification algorithm was then measured based on the number of principal

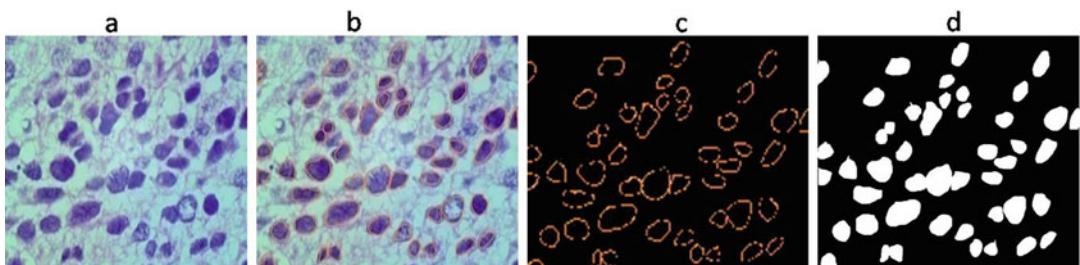


Fig. 5 Image showing the extraction of cells from the annotated ground truth. (a) is the original image, (b) is the annotated ground truth image where the boundary of the cells is marked in red, (c) is the extraction of the

boundaries of the individual cell from the annotated images, and (d) is the final binarization of the extracted cell boundary where cells are colored white and the background as black

components. This, therefore, reduced the above set of obtained feature set for the classification task.

Dividing training and test dataset: Every ML or DL classification model needs to be trained on a set of training, validation, and test data. The data is divided into train and test set initially as per choice. These training sets then use a k-fold cross-validation strategy. This helps to monitor the bias and variance trade-off in a classification model. A k-fold cross-validation could divide the entire dataset into k subsets where the training is done for the k-1 training set, leaving one subset as the validation set. The process is repeated for k times where every element gets to be in the train and test set, removing data biasness.

Classification using ML method: There are numerous ML classifiers available in the literature, and there is particularly no definite rule as to which model will fit your data best. The classification [55] was done using six classifiers, namely, decision tree, logistic regression, KNN, SVM, linear discriminant, and quadratic discriminant classifier. The best classifier was chosen based on the average performance of the classifier for the different sets of features tested during feature selection. A fivefold validation was used for the whole process.

Classification using the DL method: A DL classification method is trained by passing the input image into a series of convolution layers. An activation function is present in each layer to introduce nonlinearity in the features. Pooling layers are available for reducing the dimension of the input data and capturing maximum information. The DL

classification [58] was done using two transfer learning model of AlexNet and Vgg16. This model was trained on ImageNet dataset of natural images with 1000 classes. The model was trained to fit our two-class and four-class classification problems. Image patches were generated from the captured 10x images to feed as input to the network models. The input size of the images was 227×227 . The extracted features of these models were also compared using machine learning SVM classifier. The features extracted by VGG-16 showed a higher performance for the collected dataset.

Cell segmentation using ML and DL methods: Segmentation is the extraction of defined ROI without including the background information. The cell segmentation [59] was done using 8 ML segmentation algorithms and four deep learning semantic segmentation methods. The segmentation of the cells was both semantic and instance-based. Segmentation for ML was achieved using fuzzy C-means, entropy segmentation, k-means, adaptive thresholding, Otsu, HSV color space segmentation, YCbCr color space segmentation, and watershed method (Fig. 6). There is no training or test set involved for performing traditional segmentation, while for DL, we need to train the network using a training test set. The training is supervised and consists of a set of images with a targeted class. The DL segmentation method performed a two-class classification of the image intensity value by giving a probabilistic estimation value of a pixel belonging to the cytoplasm or

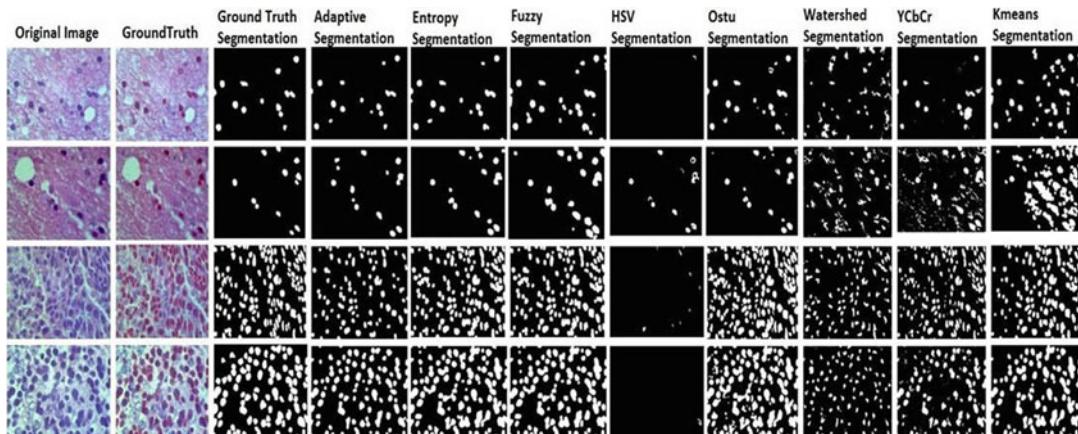


Fig. 6 Image showing the cell extraction by different segmentation methods [59]

nucleus. It consists of an encoding layer to extract valuable features from the input data, followed by a decoding layer that maps this encoded information back to the original image size. The encoding layer has a series of convolution and max pool layers for image understanding, while the decoding layer consists of convolution and up-pool layers. The training images consist of RGB images with the ground truth annotations as the targeted class image. The images were trained with an image size of 96×128 . Fractal-UNet, UNet, and SegNet are few semantic segmentation networks meant for segmentation tasks. As an output, the RGB $100 \times$ images were converted into a binary image with cells as white and the cytoplasm as black. Necessary elimination of noise was performed through morphology operation in the segmentation. Comparing the ML and DL technique of segmentation, the Fractal-UNet performed the best, while for ML methods, entropy-based segmentation performed better than all other methods.

Conclusion: The authors were, therefore, able to design an AI-based solution for the classification of CMB into the WHO grading system. A DL method depends on the architecture of the network and could be customized as per the problem and expertise of the researcher. The classification performed in the study was with a sequential DL classification network model. However, graph-based models like GoogleNet, Inception, ResNet, and DenseNet are more deep and complex and incorporate latest evaluation network architecture.

Significance of an Accurate Detection in Prognosis

Accurate detection in prognosis for medical data is highly essential not to end in under- or over-treatment. An effective prognosis can give risk-centric information to chronic syndromes [61]. The studies by [62, 63] suggest a prognosis framework for the clinicians in their decision support. An accurate biomedical diagnosis could integrate information on pathological, psychological, environmental, and behavioral factors that could channelize the treatment selection for acute

diseases. Aging factor in population is adding complexity to diagnosis and required healthcare professionals to consider numerous factors such as side effects, comorbidity, and medical interaction [64] in aged patients.

Challenges of a Clinical Observation

It is discussed that erroneous medical reports are the third-highest cause of mortality for patient care due to substandard care or poor guideline of treatment [65] [66]. has discussed a few of the important aspects of challenges faced for neurological diseases. One prime challenge discussed is the evolution of neurological disease with time, which often turns very difficult due to ambiguity at the onset and variation of defined markers that progress over time during follow-up visits. Secondly, the success of a neurological treatment is based on the early detection and intervention of the treatment. An optimized outcome is the result of an early and effective prediction and duration of treatment, which is delayed in clinical observation. Monitoring and proper follow-up of patient treatment are the next significant difficulty for an unpredictable course that requires prolonged care. Also, the diagnosis by clinician observation has variation in the treatment course and reports across various medical centers. Clinical observation is very challenging for CNS tumors and requires the experience of a neuropathologist for the appropriate report. Increase patient volume is another drawback in proper assessment under pure clinical observation. Patients from a diverse population can also have a significant complexity in the diagnosis process.

Advantages of AI-Driven Solution

An AI-driven solution can be an effective tool for precise radiological or pathological report generation. It can integrate big data information into a single decision support model for effective prediction. Moreover, an AI solution would have a faster response time than a manual testing model. It could be used to evaluate complex cases and

reports that are not possible to compute through natural intelligence. Apart from biomarkers, it would also integrate computer-based features for detection in healthcare. An AI solution could minimize Type I and Type II errors caused by manual testing. AI solutions could also be region- and application-specific and could be created with local data.

Challenges of an AI-Driven Solution

The main challenge of AI is getting access to medical data. The generation of ground truth is not possible without medical guidance and supervision. Collaboration with a medical expert is a problem as it expects dedication and time of medical collaborator, which is difficult for a medical professional to devote to such a high throughput of patients. Ethical clearance is another aspect that works as a drawback for the work, as clearance from all research collaboration needs to be acquired. The study by [67] has discussed the demerits of an AI solution. Patient data privacy and security need to be addressed for an AI solution. Training of the AI solution should be based on indigenous samples more than artificial data or benchmark dataset. Rigorous testing of the solution should be made at different clinical centers. The AI solution should be able to handle unknown problems. Overfitting of data should be handled in such applications. Equipment malfunctioning could increase surgical error, developing a single system of AI for human anatomy. Most of the study pertains to an inspection of only a single clinical task.

Scope for Industry Transformation

As stated in a market analysis [68], there is a high demand for the AI healthcare market in the world and is estimated to grow manifold by 2025. Health and economy are linked together. People in advanced countries are realizing the value of major advances in technology in healthcare with higher economic and quality of life benefits. The people in the lesser advanced countries will

benefit only if they can overcome barriers to technology implementation. Research scientists require collaboration from industries and a long time effort for the proper transition of algorithms to a product. This could lead to practical and real-life applications with proper field testing. Proper training to clinicians for the use and operation of such devices might help to mitigate such solutions [67].

Summary

The transitions of AI in neurology have already been started, as discussed in this chapter. Various studies and medical research are being made every day for refinement of prediction results and achieving a complex task. It is a fact that AI solutions could deliver better than human intelligence, and neurologist should be flexible and supportive enough to adapt to the change of AI-driven solutions. However, we could not deny that every application or innovation has its own set of disadvantages. AI will be adapted in neurology only when it will show proven results in prediction, incurring a low cost [69]. The aim of AI is never to replace medical supervision but to assist clinicians in a better understanding of their patient information. Both medical society and computer scientists should collaborate for the human good for building the transformation in neurological sciences. The transition has already started, and in a few decades, AI will soon be an integral part of a neuronal diagnosis.

References

1. Ferryman J, et al. Automated scene understanding for airport aprons. In: Zhang S, Jarvis R, editors. Advances in artificial intelligence. Berlin/Heidelberg: Springer; 2005. https://doi.org/10.1007/11589990_62.
2. Nadeem U, et al. Deep learning for scene understanding. In: Balas V, Roy S, Sharma D, Samui P, editors. Handbook of deep learning applications. Smart innovation, systems and technologies. Cham: Springer; 2019. https://doi.org/10.1007/978-3-030-11479-4_2.
3. Surendran R, Jude HD. Scene understanding using deep neural networks—objects, actions, and events: a review. In: Khanna A, Gupta D, Bhattacharyya S,

- Snasel V, Platos J, Hassanien A, editors. Advances in intelligent systems and computing. Singapore: Springer; 2020. https://doi.org/10.1007/978-981-15-1286-5_19.
4. Rong G, et al. Artificial intelligence in healthcare: review and prediction case studies. *Engineering.* 2020;6(3):291–301. <https://doi.org/10.1016/j.eng.2019.08.015>.
5. Nguyen TL, Do TTH. Artificial intelligence in healthcare: a new technology benefit for both patients and doctors. In: Proceedings of Portland international conference on management of engineering and technology (PICMET), USA. 2019. p. 1–15. <https://doi.org/10.23919/PICMET.2019.8893884>.
6. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J.* 2019;6(2): 94–8. <https://doi.org/10.7861/futurehosp.6-2-94>.
7. Staartjes VE, et al. Machine learning in neurosurgery: a global survey. *Acta Neurochir.* 2020. <https://doi.org/10.1007/s00701-020-04532-1>.
8. Senders JT, et al. Machine learning and neurosurgical outcome prediction: a systematic review. *World Neurosurg.* 2018;109:476–86.
9. Siccoli A, et al. Machine learning-based preoperative predictive analytics for lumbar spinal stenosis. *Neurosurg Focus.* 2019;46(5):E5.
10. Azimi P, et al. Use of artificial neural networks to predict surgical satisfaction in patients with lumbar spinal canal stenosis: clinical article. *J Neurosurg Spine.* 2014;20(3):300–5.
11. Senders JT, et al. Natural and artificial intelligence in neurosurgery: a systematic review. *Neurosurgery.* 2017. <https://doi.org/10.1093/neuros/nyx384>.
12. Senders JT, et al. An introduction and overview of machine learning I n neurosurgical care. *Acta Neurochir.* 2018;160(1):29–38.
13. Komura D, Ishikawa S. Machine learning methods for histopathological image analysis. *Comput Struct Biotechnol J.* 2018;16:34–42.
14. Swinburne NC, et al. Machine learning for semi-automated classification of glioblastoma, brain metastasis and central nervous system lymphoma using magnetic resonance advanced imaging. *Ann Transl Med.* 2019;7(11):232.
15. Titano JJ, et al. Automated deepneural- network surveillance of cranial images for acute neurologic events. *Nat Med.* 2018;24(9):1337–41.
16. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25(1):44–56.
17. Jalali A, et al. Deep learning for improved risk prediction in surgical outcomes. *Sci Rep.* 2010;10(1):9289. <https://doi.org/10.1038/s41598-020-62971-3>.
18. Bernert RA, et al. Artificial intelligence and suicide prevention: a systematic review of machine learning investigations. *Int J Environ Res Public Health.* 2020;17(16):5929. <https://doi.org/10.3390/ijerph17165929>.
19. Astell AJ, et al. Technology and dementia: the future is now. *Dement Geriatr Cogn Disord.* 2019;47(3):131–9. <https://doi.org/10.1159/000497800>.
20. Shahamiri SR, et al. A new autism screening system based on artificial intelligence. *Cogn Comput.* 2020;12:766–77. <https://doi.org/10.1007/s12559-020-09743-3>.
21. Abbas H, et al. Multi-modular AI approach to streamline autism diagnosis in young children. *Sci Rep.* 2020;10:5014. <https://doi.org/10.1038/s41598-020-61213-w>.
22. Khan A, Usman M. Early diagnosis of Alzheimer's disease using machine learning techniques: a review paper. In: Proceedings of 7th international joint conference on knowledge discovery, knowledge engineering and knowledge management (IC3K), Lisbon. 2015. p. 380–7.
23. Fisher CK, et al. Machine learning for comprehensive forecasting of Alzheimer's Disease progression. *Sci Rep.* 2019;9:13622. <https://doi.org/10.1038/s41598-019-49656>.
24. Kuo PJ, et al. Derivation and validation of different machine-learning models in mortality prediction of trauma in motorcycle riders: a cross-sectional retrospective study in southern Taiwan. *BMJ Open.* 2018;8:1–11.
25. Murray NM, Unberath M, Hager GD, et al. Artificial intelligence to diagnose ischemic stroke and identify large vessel occlusions: a systematic review. *J NeuroIntervent Surg.* 2020;12:156–64.
26. Yedavalli VS, et al. Artificial intelligence in stroke imaging: current and future perspectives. *Clin Imaging.* 2021;69:246–54.
27. Melo R, et al. Computer vision system with deep learning for robotic arm control. In: Proceedings of Latin American robotic symposium, 2018 Brazilian Symposium on Robotics (SBR) and 2018 Workshop on Robotics in Education (WRE), Joao Pessoa. 2018. p. 357–62. <https://doi.org/10.1109/LARS/SBR/WRE.2018.00071>.
28. Van Niftrik CHB, van der Wouden F, Staartjes VE, et al. Machine learning algorithm identifies patients at high risk for early complications after intracranial tumor surgery: registry-based cohort study. *Neurosurgery.* 2019. <https://doi.org/10.1093/neuros/nyz145>.
29. Siccoli A, et al. Machine learning-based preoperative predictive analytics for lumbar spinal stenosis. *Neurosurg Focus.* 2019;46(5):E5.
30. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol.* 2017;2. <https://doi.org/10.1136/svn-2017-000101>.
31. Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. *Z Med Phys.* 2019;29(2):102–27.
32. Bhanumurthy MY, Anne K. An automated detection and segmentation of tumor in brain MRI using artificial intelligence. In: Proceedings of 2014 IEEE international conference on computational intelligence and

- computing research, Coimbatore. 2014. p. 1–6. <https://doi.org/10.1109/ICCIC.2014.7238374>.
33. Zhu G, Jiang B, Tong L, Xie Y, Zaharchuk G, Wintermark M. Applications of deep learning to neuro-imaging techniques. *Front Neurol.* 2019;10: 869. <https://doi.org/10.3389/fneur.2019.00869>.
 34. Kong Z, et al. Automatic tissue image segmentation based on image processing and deep learning image segmentation techniques for healthcare systems. <https://doi.org/10.1155/2019/2912458>.
 35. Kerr WT, Nguyen ST, Cho AY, et al. Computer-aided diagnosis and localization of lateralized temporal lobe epilepsy using interictal FDG-PET. *Front Neurol.* 2013;4:1–14.
 36. Chiang S, Levin HS, Haneef Z. Computer-automated focus lateralization of temporal lobe epilepsy using fMRI. *J Magn Reson Imaging.* 2015;41(6):1689–94.
 37. Cohen KB, Glass B, Greiner HM, et al. Methodological issues in predicting pediatric epilepsy surgery candidates through natural language processing and machine learning. *Biomed Inform Insights.* 2016;8(8):11–8.
 38. Uddin M, Wang Y, Woodbury-Smith M. Artificial intelligence for precision medicine in neurodevelopmental disorders. *npj Digit Med.* 2019;2:112. <https://doi.org/10.1038/s41746-019-0191-0>.
 39. Dias R, Torkamani A. Artificial intelligence in clinical and genomic diagnostics. *Genome Med.* 2019;11:70. <https://doi.org/10.1186/s13073-019-0689-8>.
 40. Bozhkov L, Georgieva P, Trifonov R. Brain neural data analysis using machine learning feature selection and classification methods. In: Mladenov V, Jayne C, Iliadis L, editors. *Engineering applications of neural networks. Communications in computer and information science.* Cham: Springer; 2014. p. 459. https://doi.org/10.1007/978-3-319-11071-4_12.
 41. Vu MT, et al. A shared vision for machine learning in neuroscience. *J Neurosci.* 2018;38(7):1601–7. <https://doi.org/10.1523/JNEUROSCI.0508-17.2018>.
 42. Vieira S, et al. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications. *Neurosci Biobehav Rev.* 2017;74:58–75.
 43. Wang Z, Fey AM. SATR-DL: improving surgical skill assessment and task recognition in robot-assisted surgery with deep neural networks. In: *Proceedings of annual international conference on IEEE Eng Med Biol Soc.* 2018; 1793–1796. <https://doi.org/10.1109/EMBC.2018.8512575>.
 44. Nadeem, et al. Brain tumor analysis empowered with deep learning: a review, taxonomy, and future challenges. *Brain Sci.* 2020;10(2):118. <https://doi.org/10.3390/brainsci10020118>.
 45. Marquet G, et al. Grading glioma tumors using OWL-DL and NCI Thesaurus. In: *Proceedings of AMIA ... annual symposium proceedings. AMIA Symposium.* 2007. p. 508–12.
 46. Aneja S, Chang E, Omuro A. Applications of artificial intelligence in neuro-oncology. *Curr Opin Neurol.* 2019;32(6):850–6. <https://doi.org/10.1097/WCO.0000000000000761>.
 47. Hwang EI, Packer RJ. *Childhood brain tumors.* Elsevier; 2014.
 48. Buckner JC, et al. Central nervous system tumors. *Mayo Clin Proc.* 2007;82(10):1271–86. <https://doi.org/10.4065/82.10.1271>.
 49. Kleihues P, et al. The WHO classification of tumors of the nervous system. *J Neuropathol Exp Neurol.* 2002;61(3):215–25. <https://doi.org/10.1093/jnen/61.3.215>.
 50. Borowska A, Jóźwiak J. Medulloblastoma: molecular pathways and histopathological classification. *Arch Med Sci: AMS.* 2016;12(3):659–66. <https://doi.org/10.5114/aoms.2016.59939>.
 51. Graham ID, Lantos LP. *Greenfield's neuropathology.* CRC Press; 2015.
 52. Vinchon M, Leblond P. Medulloblastoma: clinical presentation. *Neurochirurgie.* 2019;67:23.
 53. Das D, et al. A study on MANOVA as an effective feature reduction technique in classification of childhood medulloblastoma and its subtypes. *Netw Model Anal Health Inform Bioinforma.* 2020; 9(16). <https://doi.org/10.1007/s13721-020-0221-5>.
 54. Das D, et al. Pediatric medulloblastoma— a complete study on its subtypes, characteristics and variants with regards to automated histopathological diagnosis. *Int J Appl Eng Res.* 2018;13(11):9909–15.
 55. Das D, et al. Study on contribution of biological interpretable and computer aided features towards the classification of childhood medulloblastoma cells. *J Med Syst.* 2018; 42(151). <https://doi.org/10.1007/s10916-018-1008-4>.
 56. Das D, et al. Classification of childhood medulloblastoma into W.H.O. defined multiple subtypes based on textural analysis. *J Microsc.* 2020. <https://doi.org/10.1111/jmi.12893.2020>.
 57. Das D, et al. Automated classification of childhood brain tumours based on texture feature. *Songklanakarin J Sci Technol.* 2019;41(5):1014–20.
 58. Das D, et al. Classification of childhood medulloblastoma and its subtypes using transfer learning features – a comparative study of deep convolutional neural networks. In: *Proceedings of 2020 international conference on computer, electrical & communication engineering (ICCECE), Kolkata, India.* 2020; 1–5. <https://doi.org/10.1109/ICCECE48148.2020.9223104>.
 59. Das D, Mahanta LB. On the study of childhood medulloblastoma auto cell segmentation from histopathological tissue samples. Springer. LNCS 11942. ISBN 978-3-030-34871-7.
 60. Das D, Mahanta LB. Childhood medulloblastoma microscopic images. *IEEE Dataport.* 2020. [Online]. <https://doi.org/10.21227/w0m0-mw21>. Accessed 28 Oct 2020.
 61. Vickers AJ, Basch E, Kattan MW. Against diagnosis. *Ann Intern Med.* 2008;149(3):200–3. <https://doi.org/>

- [10.7326/0003-4819-149-3-200808050-00010](https://doi.org/10.7326/0003-4819-149-3-200808050-00010). PMID: 18678847; PMCID: PMC2677291.
62. Emblem KE, Pinho MC, Zollner FG, et al. A generic support vector machine model for preoperative glioma survival associations. *Radiology*. 2015;275(1):228–34.
63. Rughani AI, Dumont TM, Lu Z, et al. Use of an artificial neural network to predict head injury outcome. *J Neurosurg*. 2010;113(3):585–90.
64. David E, et al. Acute diagnostic neurology: challenges and opportunities. *Acad Emerg Med*. 22:357. <https://doi.org/10.1111/acem.12614>.
65. Brennan TA, et al. Outcomes of medical-malpractice litigation. *N Engl J Med*. 1997;336:1680–1.
66. Jayalakshmi S, Vooturi S. Legal challenges in neurological practice. *Ann Indian Acad Neurol*. 2016;19 (Suppl 1):S3–8. <https://doi.org/10.4103/0972-2327.192888>. PMID: 27891018; PMCID: PMC5109758.
67. Panesar SS, et al. Promises and perils of artificial intelligence in neurosurgery. *Neurosurgery*. 2020;87(1):33–44. <https://doi.org/10.1093/neuroz/nyz471>.
68. Grand View Research. Artificial intelligence in healthcare market size, share & trends analysis report by component (Hardware, Software, Services), by application, by region, competitive insights, and segment forecasts, 2019–2025. 2019. <https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-healthcare-market>. Accessed Jan 2020.
69. Ganapathy K, Abdul SS, Nursetyo AA. Artificial intelligence in neurosciences: a clinician's perspective. *Neurol India*. 2018;66(4):934–9. <https://doi.org/10.4103/0028-3886.236971>.



AIM in Neurodegenerative Diseases: 120 Parkinson and Alzheimer

Joseph Davids and Hutan Ashrafiān

Contents

Introduction	1676
Artificial Intelligence for Dementia	1677
AI for Dementia Diagnosis Using Big Data	1677
Conditional Restricted Boltzmann Machines in Alzheimer's Disease	1679
Computer Vision for Dementia Patient Video Monitoring and Analysis	1679
AI and Assistive Robotic Technologies for Dementia	1679
Cognitive and Behavioral Biomarker, Facial Motion Assessment Using AI	1680
Dementia-Related Electroencephalographic Analysis and Robotic-Assisted AI	1680
Vascular Dementia and AI	1681
Convolutional Neural Nets and Model Explainability in Dementia AI Studies	1681
Artificial Intelligence in Parkinson's Disease	1681
AI in Lewy Body Dementia	1681
Motor and Gait Impairment Detection Using AI	1682
AI for Electroencephalographic Diagnosis and Prognostication in Parkinson's Disease	1683
AI for Parkinson's Disease Medical Management Drug Repurposing	1683
AI for Parkinson's Disease Surgical Management	1685

J. Davids (✉)

Imperial College London NHS Trust, London, UK

Hamlyn Centre for Robotics and Artificial Intelligence,
Department of Surgery and Cancer, Imperial College
London, London, UK

National Hospital for Neurology and Neurosurgery Queen
Square, London, UK
e-mail: j davids@ic.ac.uk

H. Ashrafiān
Department of Surgery and Cancer, Imperial College
London NHS Trust, London, UK

Institute of Global Health Innovation, Imperial College
London, London, UK

Hamlyn Centre for Robotics and Artificial Intelligence,
Department of Surgery and Cancer, Imperial College
London, London, UK

Ethical and Social Implications of AI for Parkinson's and Dementia	1686
Future In Vivo Detection and Management of Dementia and Parkinson's Using Quantum AI Systems	1686
References	1687

Abstract

Parkinson's disease and dementia are two of the most debilitating neurodegenerative disorders to ever plague humankind. They cause significant biopsychosocial and economic burden on society and affect the community and carers in particular, necessitating holistic multidisciplinary care.

The rise of artificial intelligence for medical applications in recent years, including disease prediction, diagnostics, disease progression monitoring, risk stratification, and prognostication, has also seen the development of applications for Parkinson's disease and dementia. This chapter explores the use of artificial intelligence and machine learning in terms of the diagnosis, management, and prognosis predictions for these two neurodegenerative conditions.

We discuss the medical and the surgical applications of AI for Parkinson's disease and also highlight the artificial intelligent models that have been used for various forms of dementia. The chapter begins by introducing the reader to the impacts of AI on dementia diagnosis, treatment, and prognosis, extending the discussion to dementia with Lewy body disease before tackling specific aspects of AI related to Parkinson's disease.

Keywords

AI parkinson's disease · Dementia · Vascular dementia · Frontotemporal dementia · Deep learning · Deep brain stimulation

Introduction

Dementia, and its predominant form Alzheimer's disease, remains the commonest neurodegenerative disorder; every 3 seconds someone in the

world is diagnosed with the condition [1]. Fifty million people are affected by dementia and the incidence is doubling every 20 years, due to the global aging population [1]. This entails considerable physico-biopsychosocial morbidity with damaging wide-reaching global economic consequences [1]. The burden of the disease is not just limited to the patient, but to their family and carers as well. It presents with short-term memory impairment, language, problem-solving, and high-executive dysfunction with a subsequent decline to long-term memory and neurocognitive impairment.

In the same vein, Parkinson's disease is another progressive neurodegenerative disorder affecting over 7–10 million patients globally with debilitating phenotypic changes affecting motor and cognitive pathways in the brain [2]. A classical clinical presentation is a triad-state of tremor, bradykinesia, and rigidity. In studies investigating the accuracy of clinical diagnosis using traditional methods, Parkinson's disease was a post-mortem neuropathological diagnosis with a sensitivity and specificity of 88% and 68%, respectively [3]. There is thus an argument for improving diagnostic accuracy and disease progression monitoring.

The development and revolutionary impact of artificial intelligence and machine learning has now allowed various aspects of dementia to be managed. Various chapters within this compendium of works have highlighted a myriad of approaches using machine learning and AI to tackle medical problems and we summarize the main approaches in Fig. 1.

The first part of the chapter addresses the adoption of AI for various types of dementia while the subsequent part explores Parkinson's disease and its associated Lewy body dementia that has also gained AI integration. Finally, the facets of Parkinson's disease that have benefited from artificial intelligence are discussed.

Machine Learning Cycle

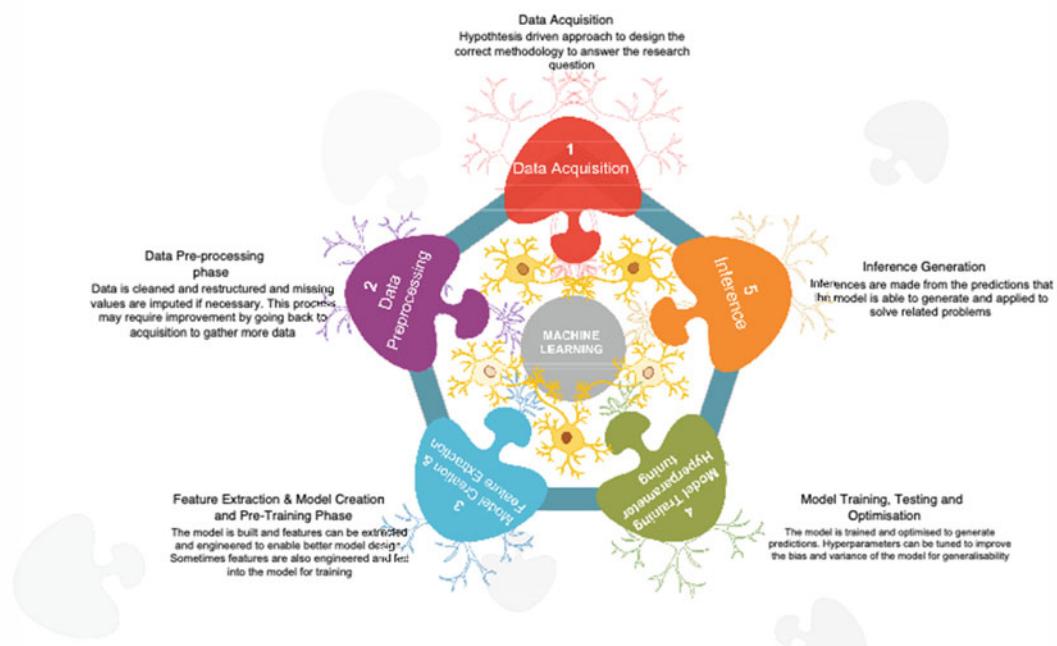


Fig. 1 A condensed variation on the machine learning cycle showing the individual aspects that are commonly used for predictive analytics

Artificial Intelligence for Dementia

Diagnosing dementia is not without its challenges. When two patients with the disease are compared, they may each present with completely different symptoms and signs of the condition including variations in their progression and treatment response. The diagnosis requires monitoring over a period of time in conjunction with imaging to correlate the clinical symptoms and signs. There are no established chemical biomarkers to aid diagnosis and as such Alzheimer's remains a clinical, neuroradiological, and post-mortem diagnosis. Computational models have been suggested to help diagnose the disease and predict its progression, and machine learning offers a chance for an unsupervised approach to feature discovery.

Results from a meta-analysis demonstrated that machine learning can be applied to neuropsychological measures with successful automatic classification as a screening tool in dementia and

as a prognostic tool identifying categories that can maximize the dementia diagnostic accuracy, thus early detection and even prevention are key [4].

AI for Dementia Diagnosis Using Big Data

Modern clinical medicine leverages impeccable detective instincts, a cooperative patient, and rapport building in a quintessential patient-doctor relationship. Dementia and its progression challenges and limits this significantly. Neurological history-taking and examination lead to the diagnosis of many neuropathological diseases, and developing approaches to augment this in dementia is therefore an area that requires considerable research investment. Attempts have also been made to incorporate machine learning in this regard. History-based Artificial Intelligent Clinical Dementia Diagnostic Systems have been

reported. In a study involving a total of 234 participants, a newly standardized and structured dementia survey was completed by members of the clinical multidisciplinary team ranging from neurologists, neuropsychiatrists, neuroradiologists, and nuclear medicine consultants [5]. The questionnaire captured brain MRI, blood tests, and neuropsychological tests as well as nuclear imaging information including Tc99m-TRODAT and Tc99m ECD SPECT. Delphi consensus style comparisons correlating history-based diagnosis of various dementia subtypes with the final diagnosis established a high degree of accuracy (76%) and identified disease severity in 86% [5].

Most machine learning algorithms rely on the need for big data to train on and make predictions. The commonest approach is usually a supervised methodology where labeled data is fed to the algorithms, which are taught to identify and learn salient properties within the data to make a prediction on diagnosis. Having a single model that is able to predict a multitude of features simultaneously is ideal, but medical datasets are often poly-dimensional and hence the single-endpoint biased predictions of the supervised approach remain suboptimal. This has led to the adoption of unsupervised artificial neural networks (ANNs) and their ability to identify latent features in a minimally biased way. ANNs offer the potential for encoding multi-modal clinical data for predictive modeling. However, clinical data is also highly sparse with missing data elements causing significant preprocessing challenges. Methods of data encoding are therefore required to overcome such limitations. However, newer methods to overcome these limitations would rely heavily on an unsupervised learning paradigm for successful implementation.

Alzheimer's disease is usually investigated using Mini Mental State Exam (MMSE) and Alzheimer's Disease Cognitive Assessment Scale (ADAS), which identify mild cognitive impairments. However, the multifactorial causes of Alzheimer's disease also contribute to its differential diagnostic heterogeneity and present a challenge for rapid diagnosis. This challenge becomes a problem that machine learning can be applied to solve. It seems reasonable that single

endpoint disease progression prediction models remain suboptimal when the disease could have variable origins and endpoints. Currently available data repositories include the Coalition Against Major Diseases (CAMD) Online Data Repository for AD (CODR-AD) [6].

Machine learning-capable mobile phone-based screening systems like DementiaTest have been built for early detection through the screening and identification of dementia phenotypes using the Diagnostic and Statistical Manual of Mental Disorders DSM V classification [7].

Frontotemporal dementia (FTD) is another area of neurodegenerative dementia affecting the frontotemporal lobe with a poor prognosis that has also seen applications in machine learning. One such work developed a reward signal-based system where an artificial intelligence algorithm (signal represented as -1 or +1) was built to retrospectively identify behavioral and language feature variants of 47 FTD patient journeys (age range 52–80) leveraging radiographic imaging and clinical time to symptom presentation [8].

Oculomics also seeks to identify early retinal changes in Alzheimer's disease. AD patients have been shown to have retinal manifestations to their disease progression with retinal amyloid plaque deposition, making it a diagnostic target and a potential biomarker to screen for in dementia [9]. As such, this has become an active area of recent scientific interest that can lend itself to potential artificial intelligence applications [9–22]. One study reported the use of highly modular machine learning techniques on UK Biobank data ($n = 52,615$; 21,547 left fundus images, and 31,041 right fundus images), utilizing methods of image quality selection and image vessel segmentation to build a classifier achieving an accuracy of 82.44 [9].

Extra-neurological and intestinal manifestations of dementia that have been considered for diagnostic biomarkers also remain a hot topic that has galvanized machine learning methods in recent decades. One such study looked at the levels of gut microbiota that were deranged because of inflammation and leveraged an unsupervised formal concept analysis model for

knowledge discovery, extraction, and management to build an AI platform for dementia diagnostics and drug discovery. It identified and implicated low-level, yet relatively abundant, fatty chain producing microbiota from the Lachnospiraceae family together with microbiota from the Enterobacteriaceae family that have previously been described as associated with AD and Parkinson's disease [23].

Conditional Restricted Boltzmann Machines in Alzheimer's Disease

Generative and energy-based models are unique unsupervised models including restricted Boltzmann machines (RBM) that draw from a sample distribution [6]. Conditional-RBMs were used by Fisher and colleagues who extracted 18-month longitudinal progression trajectories of 1909 patients with mild cognitive impairment or AD. Their unsupervised machine learning approach was conducted in two stages: (i) random synthesis of patient profiles with the same features as real patients and (ii) simulating the time-dependent evolution of these patient features [6]. The simulation aspect is a random Markov Chain Monte-Carlo-type stochastic variable prediction of the patient's disease progression that aims to capture the inherent probabilistic processes within the patient's data to forecast changes accurately. They extracted 44 covariate features including information from the individual components of the ADAS-Cog and MMSE scores, laboratory tests, and demographics. A comparison of the conditional restricted Boltzmann model with a random forest model showed similar accuracy in their results [6].

Computer Vision for Dementia Patient Video Monitoring and Analysis

Due to cognitive impairments, dementia patients are highly prone to falls. Autodetection technologies such as "SafelyYou" have been developed for fall detection [24]. In a related study, fall-prone patients were captured on 43 wall-mounted video camera systems 24 h a day, 7 days a week over a

period of 3 months in memory care facilities in the United States. The videos were reviewed on mobile device applications. The integration of computer vision is being piloted to aid the prediction of various fall contributors and thereby mitigate falls in cognitively impaired patients [25]. This presents a legal and regulatory challenge calling into question privacy and civil liberty discussions that need to be scrutinized to ensure patient safety.

AI and Assistive Robotic Technologies for Dementia

Social interaction enhances cognition in dementia patients. As such, artificially intelligent robots have been considered to augment social interaction and reduce carer burden for patients with dementia. The Care-O-Bot is a domestic mobile assistant built by the Fraunhofer Institute for Manufacturing Engineering and Automation in Germany to support activities of daily living in dementia patients. Version 4 boasts enhanced agility and a customizable modular design, and it is also economical and capable of acting as a service robot to help fetch/carry items for patients. The Care-O-Bot is also capable of evaluating a patient's mood using machine learning algorithms [1]. However, further clinical evaluation will be of interest to the scientific community.

The Robotic Assistant for Mild Cognitive Impairment Patients (RAMCIP) at home project is a European Union horizon 2020 project consisting of eight partner institutions from six countries that developed a service robot intended to aid elderly patients with mild cognitive impairment at home. This multidisciplinary process involved engaging with carers and end-users to develop the robot with the aim of monitoring for falls, medication administration, and connecting patients with family using video-conferencing [1].

Japan is leading the way in voice activated service robots and robots for dementia palliative care, such as the robotic spoon for assisting dementia patients with feeding limitations to feed themselves, and systems that aim to improve the loss of independence association with the illness [1].

Cognitive and Behavioral Biomarker, Facial Motion Assessment Using AI

The cognitive deficits in dementia range from mild depression to severe deficits, however, and symptoms may be subtle in some patients. Cognitive assessment tools such as the Montreal Cognitive Assessment have been developed and machine learning algorithms, such as logistic regression, are being used to make predictions on cognitive impairment [26].

Patients with dementia, and predominantly in Parkinson's related Lewy body dementia as discussed later, usually develop paucity of facial expression and complexion pathognomonic for neurocognitive behavioral and emotional derangements. An AI study of 121 patients presenting with cognitive impairment compared with 117 cognitively sound participants tested five deep learning models with two optimizers and built a binary classifier able to differentiate dementia from no-dementia. The facial image was expressed as a "Face AI score." The best performance was observed with the Xception model with an Adam optimizer with an overall reported respective sensitivity, specificity, accuracy, and area under the receiver operator characteristic curve of 87.31%, 94.57%, 92.56%, and 0.9717. Significant stronger correlation was also identified between Face AI score with MMSE ($r = -0.599$, $p < 0.0001$) than with chronological age ($r = 0.321$, $p < 0.0001$) [27].

Other reported approaches include using expert systems, such as SHHHH, to manage the disruptive behavioral abnormalities associated with dementia and the use of the emotional arousal and valence evaluation tasks analyzing faces for the diagnosis of cognitive impairment in dementia patients [28, 29]. One meta-analysis employing bivariate correlation and a follow-up Rutter-Gatsonis multivariate correlation HSROC curve to correct for significant correlations (47%) produced an adjusted mean specificity of 79% and sensitivity of 83% in favor of AI to diagnose for mild cognitive impairments [30].

AI-based passive radio sensing and pervasive wearable technologies have also resulted in applications for dementia in terms of monitoring

behavioral symptoms including wandering, aggression, and sleep disturbance [31, 32]. MIT's Emerald device is a low-powered radio signal emitter capable of behavioral tracking by gathering spatially mapped time-related motion data [31].

Tackling this issue from another perspective, an unsupervised nonlinear methodology using a self-organizing map was used on a multidimensional dataset to identify feature auto-clusters that helped identify the gender-determined differences in dementia diagnoses [33].

Dementia-Related Electroencephalographic Analysis and Robotic-Assisted AI

One unique study looking at the management of the neurocognitive aspects of the disease attempted to mitigate these aspects by an assessed BRAINFIT program that aims to improve the exercise ability of cognitively impaired Alzheimer's disease [34]. These patients have suboptimal exercise capability due to their cognitive deficits that lead to reduced cardiovascular fitness and cyclical independence impairment, which in turn curtails rehabilitation, making them prone to coronary artery disease insults. The team therefore developed an AI-powered robotic-assisted platform that allowed exercise training to facilitate rehabilitation and independence and reported significant reductions in EEG-discernible brain mapping of negative frontal alpha asymmetry [34]. Combined multimodality management will continue to become paramount in the management of the disease.

Deep learning and robotic microscopy were developed for dementia diagnosis to enable members of the clinical research team to monitor protein homeostasis, which led to the development of an autophagy therapeutic flux system [35]. From a multimodality imaging perspective, support vector machines extracted weight maps that were used to aid significant diagnostic performance improvements using multimodality prediction of early and late onset of Alzheimer's disease with reported accuracies of up to 96.3% [36, 37].

Vascular Dementia and AI

At the Mayo clinic, Natural Language Processing (NLP) was leveraged to reduce the heterogeneity associated with imaging reports and to examine information loss in neuroimages. It extracted information from 1000 patient CT and MRI imaging reports to identify silent brain infarcts (SBI) and white matter disease (WMD) with an acceptable interrater agreement for SBI detection, the results for the WMD and WMD grade were good ($k = 0.88$, 95% CI 0.80–0.97; $k = 0.98$, 95% CI 0.97–1.00; and Spearman = 0.985, $p < 0.001$, respectively). Patients with silent brain infarcts subsequently develop perfusion disturbance-related leukoaraiosis or small vessel and white matter disease linked to vascular dementia [38].

Convolutional Neural Nets and Model Explainability in Dementia AI Studies

Model comprehensibility and explainability remains a challenge in the field of machine learning, particularly as deep learning makes the adoption of machine learning to clinical practice difficult and problematic. The inability to identify the diagnostic pathway and thought process of a Blackbox machine learning model affects trust among healthcare professionals, which remains a problem and is a widely debated topic [39]. Circumventing this has led to the development of activation maps generated from using convolutional neural networks overlaid on volumetric scans were discussed by Marzban and colleagues [39].

Artificial Intelligence in Parkinson's Disease

Parkinson's disease (PD) characterized by tremor, bradykinesia, and gait disturbance presents a significant clinical and economic societal burden as the second most common neurodegenerative disorder in the world. A myriad of studies have looked at the pathophysiological, biochemical, and genetic determinants to derangements

associated with the dopaminergic system usually within the nigrostriatal and extrapyramidal pathways that are affected by the condition [40]. The management can either be medical or surgical, depending on the degree of neurodegeneration exhibited in the dopaminergic system and which connectivity pathways are involved, and as such careful patient selection is necessary. Several therapies have been considered and it is generally accepted, from a medical perspective, that dopamine replacement therapy can provide symptomatic treatment for a significant majority of patients for several years. The disadvantage of this is that medical therapy continues to lose its potency with increasing time as the disease progresses. The required dose increases, but the treatment benefits either remain static or usually decrease over time leading to unacceptable drug-related side effects such as motor fluctuations with significant motor and neurocognitive impairment.

Phenotypic traits in Parkinson's disease range widely from cognitive dysfunction to motor dysfunction, and in some respects autonomic dysfunction in the so-called Parkinson plus syndromes [41]. Machine learning has been suggested as an approach to aid risk stratification, patient selection, and the prediction of a variety of these effects as well as neurocognitive disease endpoints, which usually relate to Lewy body dementia. The first part of this section will therefore explore Parkinson-related Lewy body dementia and how AI is helping to diagnose and manage the disease, while the later part of this section will look at the use of artificial intelligence in various aspects of Parkinson's disease.

AI in Lewy Body Dementia

Like Alzheimer's dementia where there is pathological post-mortem identification of accumulation of amyloid plaque and Tau protein deposition in the brain, Parkinson's disease usually has an endpoint with Lewy body deposition, which also results in dementia. This section explores artificial intelligence in Parkinson's disease, but we progress our discussion from the early part of the chapter with parkinsonian

dementia and then subsequently explore the various aspects of Parkinson's disease that have benefitted from artificial intelligence methodologies. Evidence supports alpha-synuclein and amyloid beta plaque deposition in dementia associated with Parkinson's disease with three pathological subgroups [42].

Differentiating between Lewy body dementia (LBD) and Alzheimer's dementia can present significant diagnostic challenges as they have common hypometabolic profiles on single imaging modalities [43]. A metabolic reduction in the parietal association cortex in presence of preserved somatosensory and motor cortical cortex arises in both LBD and AD, meaning that ¹⁸F-fluorodeoxyglucose positron emission tomography (FDG-PET) images fail to differentiate between them. Dysfunctional metabolic activity such as occipital hypometabolism and changes in posterior cingulate cortical activity were better discriminators, with AD demonstrating hypometabolism in the primary cingulate cortex, but normal activity in LBD. Unsupervised machine learning methodologies have been applied here and include sub-profile modeling and principal component analysis. In one such study, the authors attempted to produce spatial metabolic profiles in 50 individuals each with Lewy body dementia, Alzheimer's dementia and normal cognition to help differentiate LBD from AD through radiological profiling using modalities such as MRI, ¹⁸F-FDG-PET, and dopamine transporter (DAT) single-photon emission computed tomography (DAT-SPECT). They reported that their LBD-AD discrimination profile significantly differentiated LBD from AD with comparable accuracy to that of discriminating LBD and AD from normal individuals with an area under the curve of 93.7%, sensitivity of 94.0%, and specificity of 84.0% [43].

Other areas that have been looked at with potential artificial intelligence integration include freezing of speech deficits that can arise in LBD. Questionnaires have been designed in retrospective cohorts of 666 individuals where LBD (54.0%) showed a significantly higher frequency of positive freezing of speech (all $p < 0.001$) [44].

Motor and Gait Impairment Detection Using AI

Over two centuries after the seminal description of Parkinson's disease by Drs Parkinson, Charcot, and Gowers, we still remain limited in our current approaches to understanding the condition [41]. Categorical and periodic assessments of Parkinson's disease remain episodic and clinic based. They are also heavily reliant on patient diaries to document phenotypic abnormalities that occur during the course of the disease with specific focus on the progressively debilitating motor dysfunction. Categorizing motor dysfunction using wearable sensors appears to have gained momentum over the past decade with the aim of identifying phenotypic motor-related dysfunction and its progression [41]. This section looks at artificial intelligence in monitoring motor dysfunction in Parkinson's disease.

Bradykinesia is a key symptom of Parkinson's disease that has been studied using machine learning. DeepLabCut is a computer vision system used for the analysis of bradykinesia in the form of tapping speed assessment [45].

The NeuroQWERTY index offers a web-based visualization platform and is a recent and simple validated approach to Parkinson's diagnosis of motor impairment using the natural interaction one has with a computer keyboard within an uncontrolled and unpressured domestic setting [46]. This offers an unbiased and unpressured preliminary assessment of motor dysfunction. The approach focused on finger-keyboard interaction to assess psychomotor dysfunction and leveraged machine intelligence to algorithmically validate motor dysfunction from keyboard use. Once the algorithm is installed, the data collection occurs in the background and captures the timing information at a mean temporal resolution of 3 (SD 0.28) msec, corresponding to key press and release as well as the kinematics of any keyboard input. Analysis was then performed on an encrypted remote server only assessable by the user and software administrators to maintain privacy. A reported area under the receiver operating characteristic curve [AUC] of 0.76 and

sensitivity/specificity of 0.73/0.69 is promising and could be improved further on a larger sample size [46].

Video monitoring, as mentioned in the earlier section on Alzheimer's dementia, has been employed for parkinsonian-related motor dysfunction analysis [47]. With a train-to-test split of approximately 80:20, high-speed camera videos from 208 patients were used to train the AI algorithm to identify five motor items characteristic of the disease, such as pronation-supination, finger tapping and hand movements, and enabled screening visualization. This was assigned a performance rating score of between 0 and 4 (the higher the score, the worse the motor performance) by both the human specialist and AI algorithm. Testing was performed on 28 samples to evaluate algorithm consistency with a specialist rating score. The video-based AI rating closely correlated with specialist rating in both the main ($r = 0.862$, $p = 0.000$, Pearson's correlation) and the subscale score rating ($r = 0.537$, $p = 0.000$, Pearson's correlation) [47].

One aspect of parkinsonian diagnosis is tremor, which arises in various forms: resting tremor and action tremor; kinetic and postural forms, of which resting tremor predominates in Parkinson's disease. This is an area that has also received attention from machine learning as summarized in Table 1 below [48]. Their work utilized a naïve Bayes algorithm and also applied short-term time series Fourier transforms to a dataset of 52 Parkinson's disease patients as an approach for home monitoring of parkinsonian tremor. They used features from the Unified Parkinson's Disease Rating Scale and tested it on a cost-effective embedded microcontroller platform hosting the classification algorithm and achieved 93.8% average accuracy.

Disturbances of gait and posture characterized by shuffling and freezing are another diagnostic feature for Parkinson's disease and have been explored using machine learning algorithms [49–54]. In one study the authors described a novel Ambulosono-gait-cycle-breakdown-and-freezing-detection pattern recognition algorithm to autodetect interruptions in gait-cycle and episodic freezing [55]. Aptly named Free-D, their

AI integrates a nonlinear m-dimensional phase-space data extraction method with machine learning and Monte Carlo analysis for pattern generalization.

A summary of machine learning algorithms used for motion, tremor, and gait diagnostics by body part in Parkinson's disease is presented in Table 1.

A systematic review exploring motion and gait diagnosis for Parkinson's disease identified standard machine learning algorithms leveraging supervised learning for pattern recognition and labeled data (see reference for the in-depth review) [3]. The approaches used include electromagnetic tracking, keyboard analysis, video monitoring, inertial measurement sensing multimodal, and smartphone approaches [3, 48–54, 56–62]. Algorithms use hand-crafted features, extracted from the raw data by means of different signal processing techniques.

AI for Electroencephalographic Diagnosis and Prognostication in Parkinson's Disease

Electroencephalography (EEG) is another unconventional diagnostic modality for Parkinson's disease, but studies leveraging the analytical ability of EEG have called upon it as a use-case for AI-based diagnostics in Parkinson's disease [63]. Support vector machines and linear discriminant analysis have been compared to more advanced random forest classification with default parameterization technique. Results suggested the random forest classifier showed superior performance for prognosticating the disease with an accuracy of 77.72% [63].

AI for Parkinson's Disease Medical Management Drug Repurposing

Dopaminergic dysfunction is the culminating pathophysiological endpoint arising from nigrostriatal neurodegeneration in Parkinson's disease. As such, dopamine replacement therapy forms a cornerstone for treating Parkinson's

Table 1 Review of studies using supervised algorithms. Support Vector Machine – SVM (both linear and non-linear), Support Vector Regression – SVR, Naïve Bayes – NB, Logistic Regression – LR, Artificial Neural Network – ANN (Probabilistic Neural Network – PNN, Radial Basis Function Neural Network – RBF NN, Extreme Machine Learning – EML and Dynamic Neural Networks – DNN), k-Nearest Neighbors – kNN, Linear Discriminant Analysis – LDA, Tree-based algorithms – TREE (including Decision Trees – DT, Random Forest – RF, Random Trees – RT, Ada Boost DT, C4.5 DT, BAG

DT), Hidden Markov Models – HMM, Evolutionary algorithms (EVOL), ensembles of different algorithms(ENS). PD – Parkinson’s disease; HC – Healthy controls; UPDRS – Unified Parkinson’s disease Rating Scale; H&Y – Hoen and Yahr scale; Ac – Accuracy; Se – Sensitivity; Sp – Specificity; IMU – Inertial Measurement Unit; EM tracking – Electromagnetic tracking; FtN – Finger to nose; FT – Finger tapping; HOC – Hand opening/closing; HT – Heel tapping; SIT – Sitting; HA – Hand alternating; RT – Reaction time. Adapted from a review by Belic et al. [3, 48–54, 56–62]

Goal	Type of observed motion	Body part	Instrumentation	Subjects	Algorithm	Best performance [%]		
						Sp	Se	Ac
Diagnostic	Finger tapping	Up	EM tracking	107 PD, 49 HC	EVOL	91.8	94.6	93.5
Diagnostic	Typing	Up	Keyboard	20 PD (mild), 33 HC	ENS	97	96	
Diagnostic	Arm movements at rest, waving, and walking	Up	Smartphone	21 PD (>1 year), 21 HC	ANN	95	95	95
UPDRS scoring	FT	Up	Video	13 PD (UPDRS: 0–3)	SVM			88
UPDRS scoring	Hand tremor	Up	Smartphone	52 PD	NB			97
UPDRS scoring	FT	Up	EM tracking	107 PD, 49 HC	EVOL			≥89.7
Diagnostic	Gait	Low	Force sensor	93 PD (mild and early), 73 HC	ANN	95.89	96.77	96.38
Diagnostic	Gait, Posture	Low	Smartphone	10 PD, 10 HC	RF	97.6	98.5	98.0
Diagnostic	Gait	Low	IMU	156 PD, 424 HC	kNN			85.51
FoG detection	Gait	Low	IMU	20 PD (H&Y>2)	Linear SVM	95.6	82.2	95.4
Diagnosis	Gait	Low	Camera system & force plate	23 PD (H&Y: 2), 26 HC	RF	90	96	92.6
Classification of severity of motor disorders	Unconstrained activity	All	Multimodal	19 PD, 4 non-PD	ANN	97.1	94.9	
Assessment	FtN, FT, HOC, HT, SIT, HA	All	IMU	12 PD (H&Y: 2–3)	SVM			>95
Diagnostic	Gait, Posture, FT, RT	All	Smartphone	10 PD, 10 HC	RF	96.9	96.2	
Diagnostic (PD – H&Y I)	Gait	All	IMU	27 PD (H&Y:1–3), 27 HC	SVM			94.5

disease and Levodopa (L-DOPA) remains the main anti-parkinsonian drug of choice for clinicians. L-Dopa is not without side effects, which range from daytime somnolence and episodes of sudden sleep onset (a significant side effect for drivers who must notify the licensing authority), dyskinesias, involuntary movements, and mental disturbances (usually psychosis), due to influx of dopamine into a recalibrated brain region that is used to less dopamine [2]. This list of side effects is non-exhaustive, but one drug side effect that has been looked at for drug repurposing is L-Dopa-induced dyskinesias. There are very limited medication options to counteract this including memantine and amantadine, which also have their own side effect profiles [2]. This undoubtedly increases the issue of polypharmacy. Notwithstanding the fact that patient medications may have multiple interactions with other bioactive substances that they may already be taking. As a consequence, different methods have been considered to identify other drugs that reduce the side effect burden and interactions, and to identify newer formulations that can be repurposed for L-DOPA-induced dyskinesia. In the United States, the Orphan Drug Act facilitates the granting of special status to previously available drugs for the treatment of rare disease. In silico screening approaches using AI for novel drug re-discovery and repurposing have gained momentum with an example case study being a machine learning approach leveraging natural language processing. One online repurposing tool is Project Rephetio – a browser-based direct drug repurposing tool. Another is IBM Watson, which is used for semantic similarity ranking and has been applied to help solve this problem. For an in-depth review, the reader is directed to Johnson et al. and others [2, 64].

AI for Parkinson's Disease Surgical Management

Deep brain stimulation (DBS) is a brilliant twentieth-century translational breakthrough pioneered by Banabid et al. who identified that

reversible and high-frequency simulation is akin to ablating connectivity rich basal ganglia nuclei. The effects in some circumstances were superior to ablative parkinsonian neurosurgery [65, 66].

Advances in DBS aim to counteract the deleterious symptomatic effects of progressive motor, neurocognitive, and connectivity-related dysfunction associated with the condition [67]. DBS works using precise multimodality-linked and image-guided electrode placement into intrinsic basal ganglia structures like the subthalamic nucleus, globus pallidus internus, and thalamic nuclei. Stimulating these structures enables symptomatic improvement by modulating the well-organized and distributed eloquent brain networks and anatomical tracts that encompass the motor-cognitive pathways. These electrodes are connected to a pulse-originator with a pacemaker-like activity.

The resulting post-procedural-related quality of life is usually very good to excellent. However, it is still not without risks and complications, which include electrode fracture and malfunction, infection, electrode mal-positioning, hemorrhagic strokes, seizures, etc. [68, 69]. The subthalamic nuclei, being the most approached area for DBS electrode placement, are also not without risks and side effects such as speech impairment [68–70]. The challenge is that a lot of these impairments are highly variable, and usually patient and condition dependent. It is challenging to predict which patient with speech, motor, or neurocognitive dysfunction (or a combination of these) would benefit from surgery. Current ongoing work around the world and in the UK at the National Hospital for Neurology and Neurosurgery is investigating the use of artificial intelligence and machine learning for risk stratification.

DBS electrode placement requires precision and skill for optimal intraoperative placement. Analysis of multichannel electrode recorded (MER) signals in the STN can help the multi-disciplinary operative team to identify the optimal location for lead placement. Algorithms that have been used here include support vector machines necessary for resolving taxonomy and improved

pattern recognition of deep subcortical structures such as the basal ganglia and zona incerta and thalamic nuclei. In a six-patient cohort with UKPDS of less than 6 years and a H&Y score of <4, surgical planning was achieved with a five-channel MRI frame-link and STN-recorded MER from +10 mm target to -10 mm. A 99% accuracy was reported [71].

Ethical and Social Implications of AI for Parkinson's and Dementia

Ethico-legal concerns regarding AI have been mentioned with respect to a loss of privacy and civil liberty due to computer vision and video monitoring, loss of employment, problems of cybersecurity, and data protection, as well as the potential for misuse and the need to prevent harm. Another issue relates to the ethical implications of using robots to replace carers and the interactions between robots and patients, as potential safety concerns could arise from these interactions. Moreover, there is the question of who is legally responsible if an error occurs with the robot or AI agent [72].

Future In Vivo Detection and Management of Dementia and Parkinson's Using Quantum AI Systems

The enhanced imaging resolution for diagnosing brain disease has come a long way since the early twentieth century. MRI and its variants, SPECT and PET, have all gained respective adoption in aiding the diagnosis and management of neurodegenerative and neurocognitive disease burden. However, we are still yet to gain optimal control over how we manage these conditions. Perhaps there are other salient factors and other discoveries still to be made and this can only happen through augmented approaches that allow manipulation of matter on a quantum scale. To achieve this, better imaging resolution and systems to control matter at such scales need to be designed, built, and tested for clinical

safety and effectiveness and other out-of-the-box approaches need to be considered and developed.

This area will benefit greatly from current and ongoing advances in quantum machine learning, particularly for phenotypic and disease-specific biomarker diagnostic discovery, disease progression monitoring and prediction, with the objective of curative drug discovery and also curative surgical approaches. There is evidence that in dementia, for instance, plaque accumulation has consequent neurocognitive repercussions arising from damaging neurons, which both medical and surgical advances can help to manage in future. Preventing this process may be crucial in delaying disease onset or could potentially stop patients from developing the disease in the first place. Currently, however, medical treatment does not completely solve the problem and surgical treatment is usually reserved for cases where medical treatment options have been exhausted. As DBS continues to evolve with nanoelectrodes, other new areas that may arise include quantum surgery for neurocognitive re-calibration, Lewy body removal surgery, or precision robotic micro/nano-biotechnological ubiquitin proteosome repair using tele-platforms. Controversy surrounds whether or not plaques are implicated in cognitive dysfunction [73]. However, to see pre-mortem plaques and plot the pathological course, more evolved multimodality imaging methods will also need to be developed for dynamic surgical imaging [35, 36]. The management of neurological data to integrate connectomic aspects will also evolve to be at the medic's fingertips, and this is an area where quantum machine learning will become crucial. Pervasive devices that warn of impending danger to the cognitively impaired, robotics, etc. seem more reachable than cybernetics and cognitive transfer, but it does not mean developments will not be made in these areas. The coming decades will be exciting for dementia and Parkinson's disease management and neurological disease as a whole, and the use of artificial intelligence for both medical and surgical management will lead to significantly improved patient outcomes and economic benefits for our society.

References

1. Moyle W. The promise of technology in the future of dementia care. *Nat Rev Neurol.* 2019;15(6):353–9.
2. Johnston TH, Lacoste AMB, Visanji NP, Lang AE, Fox SH, Brotchie JM. Repurposing drugs to treat L-DOPA-induced dyskinesia in Parkinson's disease. *Neuropharmacology.* 2019;147:11–27.
3. Belic M, Bobic V, Badza M, Solaja N, Duric-Jovicic M, Kostic VS. Artificial intelligence for assisting diagnostics and assessment of Parkinson's disease: a review. *Clin Neurol Neurosurg.* 2019;105442:184.
4. Battista P, Salvatore C, Berlingeri M, Cerasa A, Castiglioni I. Artificial intelligence and neuropsychological measures: the case of Alzheimer's disease. *Neurosci Biobehav Rev.* 2020;114:211–28.
5. Chiu PY, Wei CY. A history-based computerized questionnaire for the diagnosis of severity and subtypes of dementia: design and verify. *Alzheimer Dement.* 2019;15(7 Supplement):P688–P9.
6. Fisher CK, Smith AM, Walsh JR. Machine learning for comprehensive forecasting of Alzheimer's disease progression. *Sci Rep.* 2019;9(1):13622.
7. Thabtah F, Mampusti E, Peebles D, Herradura R, Varghese J. A mobile-based screening system for data analyses of early dementia traits detection. *J Med Syst.* 2020;44(1):24.
8. Brzezicki MA, Kobetic MD, Neumann S, Pennington C. Diagnostic accuracy of frontotemporal dementia. An artificial intelligence-powered study of symptoms, imaging and clinical judgement. *Adv Med Sci.* 2019;64(2):292–302.
9. Soucy JP, Chevrefils C, Sylvestre JP, Arbour JD, Rheaume MA, Beaulieu S, et al. An amyloid ligand-free optical retinal imaging method to predict cerebral amyloid pet status. *Alzheimer Dement.* 2018;14(7 Supplement):P158.
10. Chen C, Cheung C. Retinal imaging for dementia. *J Neurol Sci.* 2019;405(Supplement):19.
11. Cheung C, Chan V, Mok V, Chen C, Wong T. Potential retinal biomarkers for dementia: what is new? *Curr Opin Neurol.* 2019;32(1):82–91.
12. DeBuc DC. Identification of retinal biomarkers in Alzheimer's disease using optical coherence tomography: recent insights, challenges and opportunities. *Alzheimer Dement.* 2019;15(7 Supplement):P156–P7.
13. Dumitrescu O, Koronyo-Hamaoui M. Retinal vessel changes in cerebrovascular disease. *Curr Opin Neurol.* 2020;33(1):87–92.
14. Huang W, Luo M, Liu X, Zhang P, Ding H, Xue W, et al. Arterial spin labeling images synthesis from sMRI using unbalanced deep discriminant learning. *IEEE Trans Med Imaging.* 2019;38(10):2338–51.
15. Lau A, Mok V, Lam B, Wong A, Lee J, Lai M, et al. Automated retinal image analysis to detect white matter hyperintensities in stroke- and dementia-free healthy subjects – a cross-validation study. *Eur Stroke J.* 2018;3(1 Supplement 1):488.
16. Lau AY, Mok V, Lee J, Fan Y, Zeng J, Lam B, et al. Retinal image analytics detects white matter hyperintensities in healthy adults. *Ann Clin Transl Neurol.* 2019;6(1):98–105.
17. Nunes A, Silva G, Duque C, Januário C, Santana I, Ambrósio AF, et al. Retinal texture biomarkers may help to discriminate between Alzheimer's, Parkinson's, and healthy controls. *PLoS One.* 2019;14(6):e0218826.
18. Park CH, Lee PH, Lee SK, Chung SJ, Shin NY. The diagnostic potential of multimodal neuroimaging measures in Parkinson's disease and atypical parkinsonism. *Brain Behav.* 2020;10(11):e01808.
19. Sharafi SM, Sylvestre J-P, Chevrefils C, Soucy J-P, Beaulieu S, Pascoal TA, et al. Vascular retinal biomarkers improves the detection of the likely cerebral amyloid status from hyperspectral retinal images. *Alzheimers Dement.* 2019;5:610–7.
20. Snyder PJ, Alber J, Alt C, Bain LJ, Bouma BE, Bouwman FH, et al. Retinal imaging in Alzheimer's and neurodegenerative diseases. *Alzheimers Dement.* 2021;17(1):103–11.
21. Tian J, Smith G, Guo H, Liu B, Pan Z, Wang Z, et al. Modular machine learning for Alzheimer's disease classification from retinal vasculature. *Sci Rep.* 2021;11(1):238.
22. Wagner SK, Fu DJ, Faes L, Liu X, Huemer J, Khalid H, et al. Insights into systemic disease through retinal imaging-based oculomics. *Transl Vis Sci Technol.* 2020;9(2):6.
23. Williams C, Parmentier F, Etcheto A, Missling C, Afshar M. Levels of gut microbiota potentially regulated through anti-inflammatory effect identified as associated to response to blarcamesine (ANAVEX2-73) in Alzheimer's disease patients in 2-year interim clinical data using KEM artificial intelligence analysis. *J Prevent Alzheimers Dis.* 2019;6(Supplement 1):S98.
24. Bayen E, Nickels S, Xiong G, Jacquemot J, Agrawal P, Bayen A, et al. Real-time video detection of falls in dementia managed care: a significant reduction of time until assistance and time on the ground in fallers thanks to safely you technology. *Alzheimers Dement.* 2019;15(7 Supplement):P457.
25. Bayen E, Jacquemot J, Netscher G, Agrawal P, Noyce LT, Bayen A. Reducing the frequency and impact of falls in dementia managed care through video monitoring and incident review. *Alzheimers Dement.* 2018;14(7 Supplement):P188.
26. Kalafatis C, Modarres MH, Marefat H, Khanbagi M, Karimi H, Vahabi Z, et al. Employing artificial intelligence in the development of a self-administered, computerised cognitive assessment for the assessment of neurodegeneration. *Alzheimers Dement.* 2019;15(7 Supplement):P1355–P6.
27. Umeda-Kameyama Y, Kameyama M, Tanaka T, Son B-K, Kojima T, Fukasawa M, et al. Screening of Alzheimer's disease by facial complexion using artificial intelligence. *Aging.* 2021;13(2):1765–72.

28. Coulson JSSHHHH. An expert system for the management of clients with vocally disruptive behaviors in dementia. *Educ Gerontol.* 2000;26(4):401–8.
29. Rutkowski T, Abe MS, Koculak M, Otake-Matsuura M. Classifying mild cognitive impairment from behavioral responses in emotional arousal and valence evaluation task – AI approach for early dementia biomarker in aging societies. *Annu Int Conf IEEE Eng Med Biol Soc.* 2020;2020:5537–43. <https://doi.org/10.1109/EMBC44109.2020.9175805>. PMID: 33019233.
30. Gardner J. Artificial intelligence and machine learning algorithms for informing the diagnostic process of mild cognitive impairment and dementia. *Arch Clin Neuropsychol.* 2019;34(6):838.
31. Vahia I, Kabelac Z, Munir U, Hoti K, May R, Monette P, et al. Identification and evaluation of behavioral symptoms in dementia using passive radio sensing and machine learning. *Am J Geriatr Psychiatr.* 2019;27(3 Supplement):S126–S7.
32. Thabtah F, Peebles D, Retzler J, Hathurusingha C. Dementia medical screening using mobile applications: a systematic review with a new mapping model. *J Biomed Inform.* 2020;103573:111.
33. Grossi E, Massini G, Buscema M, Savare R, Maurelli G. Two different Alzheimer diseases in men and women: clues from advanced neural networks and artificial intelligence. *Gend Med.* 2005;2(2):106–17.
34. Shin J, Park H, Park C, Hwang J, You SH. Effects of artificial intelligence (AI) based integrated robotic-assisted gait, music, and light Brain fitness training (BRAIN-FIT) on electroencephalography (EEG) brain mapping of frontal alpha asymmetry (FA) and associated psychological behaviors in anxiety and depression. *IBRO Rep.* 2019;6(Supplement):S288.
35. Finkbeiner S. Harnessing human brain cell models with robotics and deep learning to discover causes and treatments for neurodegenerative disease. *Acta Physiol.* 2019;227(Supplement 719):17–8.
36. Liu X, Chen K, Wu T, Weidman D, Lure F, Li J. Use of multimodality imaging and artificial intelligence for diagnosis and prognosis of early stages of Alzheimer's disease. *Transl Res.* 2018;194:56–67.
37. Wee C-Y, Yap P-T, Zhang D, Denny K, Browndyke JN, Potter GG, et al. Identification of MCI individuals using structural and functional connectivity networks. *NeuroImage.* 2012;59(3):2045–56.
38. Leung L, Fu S, Nelson J, Kallmeh D, Luetmer P, Liu H, et al. Abstract 135: examining the information loss between neuroimages and neuroimaging reports for detection of silent brain infarcts and white matter disease using artificial intelligence technologies. *Stroke.* 2020;51(Suppl_1):A135.
39. Marzban EN, Teipel SJ, Buerger K, Fließbach K, Heneka MT, Kilimann I, et al. Explainable convolutional networks and multimodal imaging data: the next step towards using artificial intelligence as diagnostic tool for early detection of Alzheimer's disease. *Alzheimers Dement.* 2019;15(7 Supplement):P1083–P4.
40. McLeod JG. Pathophysiology of Parkinson's disease. *Aust New Zeal J Med.* 1971;1:19–23. <https://doi.org/10.1111/j.1445-5994.1971.tb02561x>.
41. Dorsey ER, Omberg L, Waddell E, Adams JL, Adams R, Ali MR, et al. Deep phenotyping of Parkinson's disease. *J Parkinsons Dis.* 2020;10(3):855–73.
42. Kotzbauer PT, Cairns NJ, Campbell MC, Willis AW, Racette BA, Tabbal SD, et al. Pathologic accumulation of α -Synuclein and A β in Parkinson disease patients with dementia. *Arch Neurol.* 2012;69(10):1326–31.
43. Iizuka T, Kameyama M. Spatial metabolic profiles to discriminate dementia with Lewy bodies from Alzheimer disease. *J Neurol.* 2020;267(7):1960–9.
44. Chiu PY, Hung GU, Wei CY, Tzeng RC, Pai MC. Freezing of speech single questionnaire as a screening tool for cognitive dysfunction in patients with dementia with Lewy bodies. *Front Aging Neurosci.* 2020;12:65.
45. Williams S, Zhao Z, Hafeez A, Wong DC, Relton SD, Fang H, et al. The discerning eye of computer vision: can it measure Parkinson's finger tap bradykinesia? *J Neurol Sci.* 2020;117003:416.
46. Arroyo-Gallego T, Ledesma-Carbayo MJ, Butterworth I, Matarazzo M, Montero-Escribano P, Puertas-Martin V, et al. Detecting motor impairment in early Parkinson's disease via natural typing interaction with keyboards: validation of the neuroQWERTY approach in an uncontrolled at-home setting. *J Med Internet Res.* 2018;20(3):e89.
47. Shen B, Peng F, Xie Y, Chen Y, Tang H, Lin S, et al. A pilot study to evaluate the severity of motor dysfunction in patients with Parkinson's disease based on AI non-wearable motion capture of video analysis. *Mov Disord Clin Pract.* 2019;6(Supplement 1):S73–S4.
48. Bazgir O, Habibi SH, Palma L, Pierleoni P, Nafees S. A classification system for assessment and home monitoring of tremor in patients with Parkinson's disease. *J Med Signals Sens.* 2018;8:65–72.
49. Arora S, Venkataraman V, Zhan A, Donohue S, Biglan KM, Dorsey ER, Little MA. Detecting and monitoring the symptoms of Parkinson's disease using smartphones: a pilot study. *Parkinsonism Relat Disord.* 2015;21:650–3.
50. Wahid F, Begg RK, Hass CJ, Halgamuge S, Ackland DC. Classification of Parkinson's disease gait using spatial-temporal gait features. *IEEE J Biomed Health Inform.* 2015;19:1794–802.
51. Ahlrichs C, Samà A, Lawo M, et al. Detecting freezing of gait with a tri-axial accelerometer in Parkinson's disease patients. *Med Biol Eng Comput.* 2016;54:223–33.
52. Cuzzolin F, Sapienza M, Esser P, Saha S, Franssen MM, Collett J, Dawes H. Metric learning for Parkinsonian identification from IMU gait measurements. *Gait Posture.* 2017;54:127–32.
53. Arora S, Venkataraman V, Donohue S, Biglan KM, Dorsey ER, Little MA. High accuracy discrimination of Parkinson's disease participants from healthy

- controls using smartphones. In: Proceedings of the ICASSP, IEEE international conference on acoustics, speech and signal processing – proceedings, 2014.
- 54. Zeng W, Liu F, Wang Q, Wang Y, Ma L, Zhang Y. Parkinson's disease classification using gait analysis via deterministic learning. *Neurosci Lett.* 2016;633:268–78.
 - 55. Chomiak T, Xian W, Pei Z, Hu B. A novel single-sensor-based method for the detection of gait-cycle breakdown and freezing of gait in Parkinson's disease. *J Neural Transm.* 2019;126(8):1029–36.
 - 56. Adams WR. High-accuracy detection of early Parkinson's disease using multiple characteristics of finger movement while typing. *PLoS One.* 2017;12: e0188226.
 - 57. Fraiwan L, Khnouf R, Mashagbeh AR. Parkinsons disease hand tremor detection system for mobile application. *J Med Eng Technol.* 2016;40:127–34.
 - 58. Khan T, Nyholm D, Westin J, Dougherty M. A computer vision framework for finger-tapping evaluation in Parkinson's disease. *Artif Intell Med.* 2014;60(1):27–40.
 - 59. Gao C, Smith S, Lones M, Jamieson S, Alty J, Cosgrove J, Zhang P, Liu J, Chen Y, Du J, et al. Objective assessment of bradykinesia in Parkinson's disease using evolutionary algorithms: clinical validation. *Transl Neurodegener.* 2018;7:18.
 - 60. Roy S, Cole BT, Gilmore LD, De Luca CJ, Thomas CA, Saint-Hilaire MM, Nawab SH. High-resolution tracking of motor disorders in Parkinson's disease during unconstrained activity. *Mov Disord.* 2013;28(8):1080–7.
 - 61. Caramia C, Torricelli D, Schmid M, Munoz-Gonzalez A, Gonzalez-Vargas J, Grandas F, Pons JL. IMU-based classification of Parkinson's disease from gait: a sensitivity analysis on sensor location and feature selection. *IEEE J Biomed Health Inform.* 2018;22(6):1765–74.
 - 62. Patel S, Lorincz K, Hughes R, Huggins N, Growdon J, Standaert D, Akay M, Dy J, Welsh M, Bonato P. Monitoring motor fluctuations in patients with Parkinson's disease using wearable sensors. *IEEE Trans Inf Technol Biomed.* 2009;13:864–73.
 - 63. Soria-Frisch A, Kroupi E, Castellano M, Ibanez D, Montplaisir J, Jean-Francois G, et al. Comparison of EEG-based classifier performance for PD prodromal analysis. *Neurodegener Dis.* 2017;17(Supplement 1):1570.
 - 64. Alford SH, Visanji NP, Lacoste AMB, Madan P, Buleje I, Han Y, et al. Using artificial intelligence and realworld data to identify drugs to repurpose for Parkinson's disease. *Pharmacoepidemiol Drug Saf.* 2019;28(Supplement 2):131.
 - 65. Neumann W-J, Turner RS, Blankertz B, Mitchell T, Kuhn AA, Richardson RM. Toward electrophysiology-based intelligent adaptive deep brain stimulation for movement disorders. *Neurotherapeutics.* 2019;16(1):105–18.
 - 66. Rowland NC, Sammartino F, Lozano AM. Advances in surgery for movement disorders. *Mov Disord.* 2017;32(1):5–10.
 - 67. Gratiwicke J, Zrinzo L, Kahan J, et al. Bilateral deep brain stimulation of the nucleus basalis of Meynert for Parkinson disease dementia: a randomized clinical trial. *JAMA Neurol.* 2018;75(2):169–78. <https://doi.org/10.1001/jamaneurol20173762>.
 - 68. Limousin PKP, Pollak P, Benazzouz A, Ardouin C, Hoffmann D, Benabid A-L. Electrical stimulation of the subthalamic nucleus in advanced Parkinson's disease. *N Engl J Med.* 1998;339:1105–11.
 - 69. Hariz MI. Complications of deep brain stimulation surgery. *Mov Disord.* 2002;17:S162–6. <https://doi.org/10.1002/mds10159>.
 - 70. Phokaewvarangkul O, Boonpang K, Bhidayasiri R. Subthalamic deep brain stimulation aggravates speech problems in Parkinson's disease: objective and subjective analysis of the influence of stimulation frequency and electrode contact location. *Parkinsonism Relat Disord.* 2019;66:110–6. <https://doi.org/10.1016/j.parkreldis201907020>. Epub 2019 July 16 PMID: 31327627.
 - 71. Rama Raju V. Effectiveness of lead position with microelectrode recording based support vector machine for characterizing the sub-cortical-structures via deep brain stimulus in Parkinson's disease. *Mov Disord.* 2020;35(Suppl 1):S370.
 - 72. Fiske A, Henningsen P, Buyx A. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *J Med Internet Res.* 2019;21(5):e13216.
 - 73. Melzer TRSM, Keenan RJ, Myall DJ, MacAskill MR, Pitcher TL, Livingston L, Grenfell S, Horne K-L, Young BN, Pascoe MJ, Almuqbel MM, Wang J, Marsh SH, Miller DH, Dalrymple-Alford JC, Anderson TJ. Beta amyloid deposition is not associated with cognitive impairment in Parkinson's disease. *Front Neurol.* 2019;10:391. <https://doi.org/10.3389/fneur201900391>.



AIM in Amyotrophic Lateral Sclerosis 121

Meysam Ahangaran and Adriano Chiò

Contents

Introduction	1692
Review	1692
Clinical Trial Analysis of ALS Disease	1693
PRO-ACT Dataset	1693
Study of ALS with Machine Learning Approach	1694
Experimental Results	1697
Conclusion	1699
References	1702

Abstract

Amyotrophic lateral sclerosis (ALS) is a relentlessly progressive neurodegenerative disease of motor neurons with substantial heterogeneity in its clinical presentation. Survival ranges from

3 to 5 years after symptom onset depending on genetic, geographic, and phenotypic factors. Despite tireless research efforts, the cause of ALS remains unknown, and therapy development efforts are confounded by the lack of accurate prognosis markers. Artificial intelligence (AI) with machine learning (ML) methods offers unprecedented opportunities to construct accurate prognosis and diagnostic models. Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) dataset is the largest public ALS clinical trial data that various researches were accomplished on this framework with ML approach. In this chapter, we study the causal analysis of ALS with AI perspective, and also ML methods for predicting the evolution of ALS will be explained on the PRO-ACT dataset. This study reveals that AI-based longitudinal study of ALS by

Supplementary Information: The online version of this chapter (https://doi.org/10.1007/978-3-030-64573-1_252) contains supplementary material, which is available to authorized users.

M. Ahangaran (✉)
Computer Engineering – Artificial Intelligence, Iran
University of Science and Technology, Tehran, Iran

Mazandaran University of Science and Technology, Babol,
Iran

A. Chiò
‘Rita Levi Montalcini’ Department of Neuroscience,
University of Turin, Turin, Italy
e-mail: adriano.chio@unito.it

considering clinical, genetic, and imaging factors could bring new insight to the core etiology of ALS.

Keywords

Amyotrophic lateral sclerosis · Prognosis · Causal discovery · Longitudinal dataset · Machine learning · Information theory

Introduction

Amyotrophic lateral sclerosis (ALS) is a progressive neurodegenerative disease with predominant motor system involvement. The incidence of ALS in Europe varies between 2 and 3 cases per 100,000 individuals [1], and its prevalence is between 5 and 8 cases per 100,000 [2]. With the ever-increasing interest in artificial intelligence (AI), a large number of research papers have been recently published using machine learning (ML) techniques and predictive modeling in ALS [3, 4]. Because of the heterogeneity in ALS, the clinical progression of the disease has been difficult to assess; consequently, developing an appropriate treatment for ALS becomes challenging. The progression of ALS is generally measured by the ALS Functional Rating Scale-Revised (ALSFRS-R), which consists of 12 functional items [5]. The rate of ALS progression varies in patients and finally leads to death after about 3 years; however, various parameters, such as age of onset, clinical phenotype, geographic factors, and genetics, strongly influence survival time of the ALS patients [6]. Genetic profile of ALS provides an important layer of heterogeneity. However, specific genotype such as mutations of four genes – *C9ORF72*, *TARDBP*, *SOD1*, and *FUS* – have been associated with clinical presentation of ALS [7].

Study of clinical longitudinal dataset of the ALS with AI approach could shed new insight on the biology of its causation. Providing accurate prediction and survival estimates in ALS is challenging, because they are influenced by a wide range of demographic, genetic, and clinical factors. Typical stratifications of ALS patients

include *sporadic* vs. *familial*, *bulbar* vs. *spinal*, and *ALS-FTD* vs. *ALS* with no cognitive impairment [8]. Forecasting ALS progression is commonly based on statistical approaches and ML models. Most prognostic models have used clinical features to predict the disease progression, but recent studies enriched the clinical data with imaging measures [9, 10]. Prediction of ALS progression is typically addressed either by a classification or by a regression problem based on functional decline of the disease which is determined by the ALSFRS-R score. The Neurological Clinical Research Institute (NCRI) at Massachusetts General Hospital (MGH) constructed a Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) clinical trial database that enables greater understanding of the phenotype and biology of ALS [11]. There are various studies which proposed prognosis models based on PRO-ACT data on the basis of ML methods [12–18]. In this chapter, we will study the causal analysis of ALS disease based on longitudinal study of PRO-ACT dataset, and the prognosis modeling of ALS will be explained in order to predict disease progression from an AI perspective.

Review

ALS is a fatal neurodegenerative disorder with heterogeneity in its clinical presentation that is specified by the progressive impairment of motor functions including speech, swallowing, limb, and respiration [7]. ALS is characterized by the degeneration of both upper motor neurons (from the cortex to the brainstem and the spinal cord) and lower motor neurons (from the brainstem or spinal cord to muscle) [1]. The cause of ALS is unknown, although in 10% of individuals, ALS is caused by genetic defects in more than 30 genes [6, 19]. Because of heterogeneity in the ALS behavior, the prediction of disease progression is a challenging task. The progression of ALS is most commonly assessed by the ALS Functional Rating Scale (ALSFRS) that contains 10 functional items, and the new revised version of ALSFRS, namely, ALSFRS-R, includes 12 functional items: *speech*, *salivation*, *swallowing*,

handwriting, cutting food, dressing and hygiene, turning in bed, walking, climbing stairs, dyspnea, orthopnea, and respiratory insufficiency [5, 20].

The scores of all 12 items range between 0 (worst) and 4 (best); therefore, the overall value of ALSFRS ranges between 0 and 40, and ALSFRS-R ranges between 0 and 48.

Study of clinical datasets of the ALS with AI approach can help neuroscientists for better understanding the behavior of the disease. The disease course ranges from under a year to over 10 years, and the patients usually die between 3 and 5 years from disease onset [21]. The more heterogeneous the disease, the more difficult to predict its progression, and thereby to study the effect of a potential therapy will be a challenging task. Using a large clinical longitudinal dataset of patients with machine learning and computational methods could be a suitable approach to address the variability of ALS disease progression. This approach has been successfully applied for understanding complex diseases such as multiple sclerosis (MS), Alzheimer's, and Parkinson's [22, 23].

Clinical Trial Analysis of ALS Disease

Analyzing the behavior of a disease on a clinical trial dataset with the help of machine learning algorithms could open a door to new insights into the disease mechanisms. Prize4Life and Neurological Clinical Research Institute at Massachusetts General Hospital created the PRO-ACT dataset that includes information from over 10,000 ALS patients [11]. This dataset consists of more than 10 million longitudinally collected data points which include demographic, family histories, and clinical and laboratory data. In a crowdsourcing analysis of ALS on the PRO-ACT dataset, 3 months of trial information of the patients were given to the solvers in order to predict the disease progression over the subsequent 9 months [18]. In this challenge, the solvers used 12 months of longitudinal data for training their algorithm, and then their algorithm was evaluated on a separate test dataset not available for the training of the algorithm. The progression rate of the disease is assessed by the slope of ALSFRS

changes which is defined as Eq. 1, where m_{first} and m_{last} indicate the first and last month of prediction, respectively:

$$\text{Slope} = \frac{\text{ALSF}RS(m_{\text{last}}) - \text{ALSF}RS(m_{\text{first}})}{m_{\text{last}} - m_{\text{first}}} \quad (1)$$

The performance of methods in this challenge was measured by root-mean-square deviation (RMSD) of the predicted ALSFRS slope. The best method in this study was Bayesian tree with RMSD value of 0.544, and linear regression had the largest RMSD value of 1.30. Some features such as time from onset, age, forced vital capacity (FVC), site of onset, gender, weight, and uric acid concentration in blood have been formerly reported to predict ALS progression. This study revealed that three features – *phosphorus*, *creatinine*, and *pulse* – can be considered as predictive features to predict ALS progression which were not previously reported. Specifically, the study showed a high correlation between changes in creatinine and ALSFRS score in patients.

The great variability in ALS progression hinders to specify the effect of a given treatment, increasing the cost of clinical trials. The ability to more precisely predict the disease progression would reduce the number of patients needed for identifying the effect of a potential disease-modifying therapy. The amount of trial size reduction by using two best algorithms of this challenge was up to 20.4%, with a significant reduction in cost [18].

PRO-ACT Dataset

The PRO-ACT platform is the largest clinical longitudinal dataset, which has provided an opportunity to the study of the ALS disease with data mining and machine learning tools [24]. This dataset includes temporal clinical information from over 10,000 ALS patients, which has been collected regularly for every patient. This dataset includes the databases of 17 different clinical studies with an average duration of 12 months over the last 20 years

between 1990 and 2010 performed in ALS. The PRO-ACT dataset consists of 86 dynamic features which change over time, and the progression of the disease in each time point is assessed by ALSFRS score. The information of the patients in the dataset are divided into 13 files: adverse events, ALSFRS, subject ALS history, concomitant medication use, death report, demographics, family history, forced vital capacity (FVC), laboratory data, Riluzole use, slow vital capacity (SVC), treatment group, and vital sign. As shown in Table 1, all dynamic features of the PRO-ACT dataset are classified into 16 classes so that each class includes 1 or more features. Average values of baseline characteristics of individuals are summarized in Table 2.

Table 1 All dynamic features of the PRO-ACT dataset categorized into 16 classes in which every class contains 1 or more features

Classes	Number of features	Features names
Vital signs	6	<i>Pulse, respiratory rate, temperature, weight, blood pressure diastolic (BPD), blood pressure systolic</i>
Forced vital capacity	1	<i>FVC</i>
Slow vital capacity	1	<i>SVC</i>
Urine	17	<i>PH, protein, specific gravity, glucose, WBC, leukocyte esterase, blood, RBC, casts, ketones, appearance, color, bacteria, mucus, albumin, uric acid crystals, calcium oxalate crystals</i>
Blood proteins	2	<i>Albumin, protein</i>
Electrolytes	6	<i>Sodium, potassium, bicarbonate, chloride, anion gap, magnesium</i>
Kidney tests	3	<i>Blood urea nitrogen, uric acid, creatinine</i>
Liver tests	5	<i>ALP, ALT, GGT, AST, bilirubin total</i>
Complete blood count	20	<i>WBC, neutrophils, absolute neutrophil count, band neutrophils, absolute band neutrophil count, lymphocytes, absolute lymphocyte count, monocytes, absolute monocyte count (AMC), eosinophils, absolute eosinophil count (AEC), basophils, absolute basophil count, RBC, hemoglobin, hematocrit, mean corpuscular hemoglobin, mean corpuscular hemoglobin concentration, mean corpuscular volume, platelets</i>
Heart disease	4	<i>CK, triglycerides, total cholesterol, lactate dehydrogenase</i>
Blood sugar	2	<i>Glucose, HbA1c</i>
Mineral balance	2	<i>Calcium, phosphorus</i>
Immune response	8	<i>Immunoglobulin A, immunoglobulin G, immunoglobulin M, gamma globulin, alpha 1 globulin, alpha 2 globulin, beta globulin, albumin globulin ratio</i>
Hormones	4	<i>TSH, free T3, free T4, beta HCG</i>
Coagulation measures	2	<i>Prothrombin time, international normalized ratio (INR)</i>
Amylase	3	<i>Amylase, salivary amylase, pancreatic amylase</i>

Study of ALS with Machine Learning Approach

Machine learning is a rapidly evolving branch of AI and applied mathematics which aims to develop computer software that can learn autonomously. Machine learning encompasses two main approaches: *supervised* and *unsupervised* learning. The goal of unsupervised learning is to construct a model that describes the data in the absence of output, but supervised learning focuses on mapping inputs with outputs using training samples [25]. Classification task in ALS is to link clinical features, pathological characteristics, etc. to diagnostic classes of the disease such as “ALS” or “healthy” [4, 26, 27]. Regression methods aim to infer a real-valued function from

Table 2 Baseline characteristics of the PRO-ACT dataset and their average values

Characteristics	Mean \pm SD
Age (year)	56.2 \pm 11.8
Percentage of females	40
Percentage of white	94.9
Disease duration (month)	23 \pm 17
Time from symptom onset to diagnosis (month)	11.6 \pm 9.2
Site of onset, percentage of limb onset	76
Vital capacity, percentage of predicted normal	86 \pm 24
ALSFRS score	29.59 \pm 5.84
ALSFRS-R score	38.37 \pm 5.22
Body mass index at baseline	25.4 \pm 4.4
Percentage of Riluzole use	77.5

input data, which, in the context of ALS, could be used for predicting the progression of ALS based on clinical observations [28, 29]. Regression task can design a prognosis model from longitudinal data/set of ALS that is able to predict motor decline on the basis of clinical characteristics.

Causal discovery of diseases is a suitable approach for improving the design of clinical trials, finding new therapies, and establishing a prognosis model of such illnesses, and probabilistic causality is a common method for diagnosis and prognosis in biomedical applications. In general, the cause of ALS is unknown; however, in 10% of cases, ALS is caused by a genetic defect [19]. Discovering causal factors of ALS helps physicians to find new therapies and to establish a prognosis model for the disease. Since ALS is a chronic disorder with a heterogeneous clinical course, understanding the behavior of the disease and predicting its progression is a challenging task for neuroscientists.

In a study, the authors introduced a novel learning model named PCDSD (probabilistic causal discovery in sequential dataset) for constructing a causal graph from sequential dataset (Fig. 1). They applied the PCDSD model to the PRO-ACT longitudinal dataset for predicting the progression of ALS disease [16]. In this project, a longitudinal dataset $S = \{s_1, s_2, \dots, s_N\}$ of ALS patients from PRO-ACT dataset with n-dimensional feature space $F = \{f_1, f_2, \dots, f_n\}$ is given to the model.

Every patient s_i from S contains the clinical trial information of the corresponding patient in k consecutive time points in the form of $e_{i1}, e_{i2}, \dots, e_{ik}$.

Each event e_{ij} indicates the state of the patient s_i at time point t_j that includes the values of all 86 dynamic features of Table 1 and the value of ALSFRS score that shows the progression of the disease at time point t_j . First, the values of all features changes at all consecutive events for all patients are calculated; for every patient s_i at time point t_j , the change value of feature f_l is specified relative to its value at time point t_{j-1} as Δf_l^{ij} , and if the change value Δf_l^{ij} is greater than a threshold value τ_{f_l} , then the feature f_l is considered as a modified feature in event e_{ij} . We define Ω_{ij} as a set of all modified features for patient s_i at time point t_j according to the following definition:

$$\Omega_{ij} = \{ f_l \in F | \Delta f_l^{ij} > \tau_{f_l} \text{ in sequences}_i \text{ at time point } t_j \} \quad (2)$$

In the next step, the causal features dependency (CFD) matrix M of the PRO-ACT dataset with dimension $n \times n$ is created on the basis of Ω sets of the whole patients in all time points. Every matrix index $m_{c,e}$ of M consists of the change values Δf_c of cause feature f_c and the change values Δf_e of effect feature f_e , which indicates the probabilistic causal edge $f_c \rightarrow f_e$ with conditional probability $P(\Delta f_e | \Delta f_c)$. Therefore, the rows and columns of matrix M show the cause and effect features of the causal edges, respectively. Then, the conditional probability density (CPD) function $Z(f_c, f_e, d_{ce})$ is estimated for every causal edge $f_c \rightarrow f_e$ corresponding to the matrix index $m_{c,e}$ by kernel density estimation (KDE) method, where d_{ce} is the density value at point $(\Delta f_c, \Delta f_e)$. Here, we have n^2 density functions which their causal tendency (CT) should be determined according to Eq. 3:

$$CT_{(f_c \rightarrow f_e)} = \log \frac{p(\Delta f_e | \Delta f_c)}{p(\Delta f_e | \Delta f_c)} \quad (3)$$

The higher the value of the causal tendency, the more degree of certainty of the corresponding causal relationship. The numerator in Eq. 3

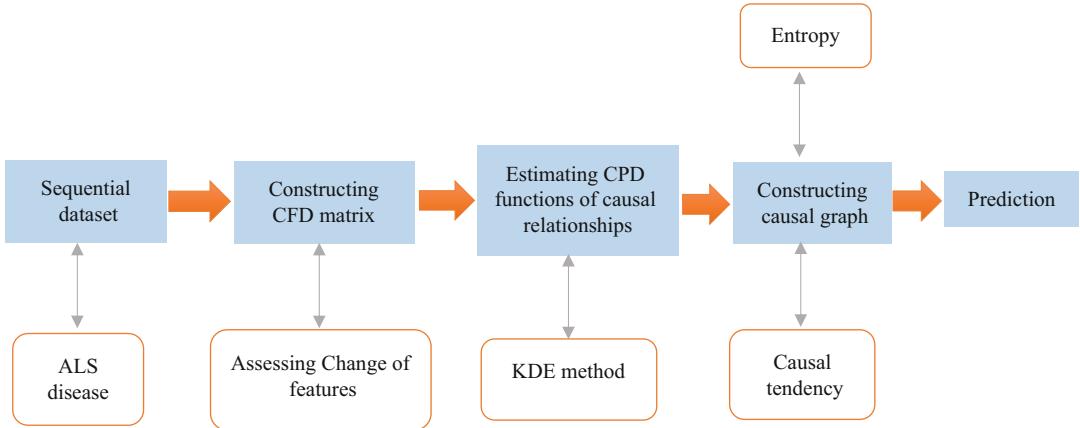


Fig. 1 Diagram of the PCDSD model for causal analysis of ALS disease on the PRO-ACT dataset

indicates the probability of changing in cause feature f_c that leads to the changes of effect feature f_e , and the denominator is the probability of changes of feature f_c which did not cause to changes of feature f_e (Eqs. 4 and 5):

$$P(\Delta f_e | \Delta f_c) \equiv \forall_{p \in \text{sequences}} \forall_{q \in \text{time points}} \quad (4)$$

$$(f_c \in \Omega_{pq} \wedge f_e \in \Omega_{p,q+1})$$

$$P(\Delta \bar{f}_e | \Delta f_c) \equiv \forall_{p \in \text{sequences}} \forall_{q \in \text{time points}} \quad (5)$$

$$(f_c \in \Omega_{pq} \wedge f_e \notin \Omega_{p,q+1})$$

In the next step, the degree of uncertainty for every CPD function $Z(f_c, f_e, d_{ce})$ corresponding to the causal edge $f_c \rightarrow f_e$ is determined by Shannon entropy measure according to the following formula [30]:

$$H_{(f_c \rightarrow f_e)} = \sum_{(\Delta f_c, \Delta f_e) \in Z} -p_{ce} \log p_{ce} \quad (6)$$

where p_{ce} indicates the conditional probability $P(\Delta f_e | \Delta f_c)$ at the corresponding point of density function. Density functions with the low entropy values show that the corresponding causal edge has a high degree of certainty so that the function has a high-density value in a small area of points. Only causal edges with the causal tendency greater than threshold value δ and entropy value less than threshold value ε are selected for constructing the causal graph G . The graph G contains different paths called *causal chains*,

and therefore every path of the causal graph could be a potential causal chain. A probabilistic causal chain $f_1 \rightarrow f_2 \rightarrow f_3 \rightarrow f_4 \rightarrow \dots$ indicates the causal dependencies between consecutive features in which every change of feature f_i leads to the change in feature f_{i+1} with high probability.

We could pursue the progression of ALS disease by tracking its features changes with the help of the set of causal chains Π . In general, searching all paths of a causal graph in order to extract most probable causal chains is an impractical task, because this procedure has exponential running time on the number of graph nodes. However, reducing the size of search space by a suitable heuristic could improve the time complexity of the algorithm significantly. In the worst case, discovering the set of most probable causal chains in a causal graph with n nodes has a time complexity of $O(n!)$, because the number of permutations of k nodes for constructing a causal chain with the length of k is $k!$, and hence the running time of the algorithm, $T(n)$, is calculated as follows:

$$T(n) = \sum_{k=1}^n k! = 1! + 2! + \dots + n! = O(n!) \quad (7)$$

In a study, the authors introduced a novel heuristic for discovering the set of most probable causal chains from a causal graph, which has a polynomial time complexity on the number of graph nodes [17]. The proposed algorithm is

based on row- and column-major processing of the CFD matrix on the causal graph G . In the *row-major* approach, the matrix M is processed row to row in order to create the causal chain until reaching to a repeated feature (row). Actually, in this approach, we move forward across causal chains in the causal graph from cause features to effect features, and the causal chains are constructed from beginning to the end (Fig. 2). In the *column-major* approach, the matrix M is processed column to column for creating causal chains until reaching to a repeated feature (column). Therefore, in this approach, we move backward across causal chains in G from effect features to cause features, and the causal chains will be created reversely from end to the beginning (Fig. 3).

Selection of the causal edges is carried out based on the causal tendency-entropy ratio ($CTER$) measure. $CTER$ is a quantity that combines both causal

tendency and entropy measures which is obtained by the following definition:

$$CTER_{(f_c \rightarrow f_e)} = \frac{CT_{(f_c \rightarrow f_e)}}{H_{(f_c \rightarrow f_e)}} \quad (8)$$

Large values of $CTER$ reveal that corresponding causal edge has more accuracy and certainty, because the large value of CT and small value of entropy indicate the high confidence level of the corresponding causal edge.

Experimental Results

In this section, the PCDSD model is applied to the PRO-ACT dataset related to ALS disease in order to discover the causal relationships between disease factors. In the training phase, trial information of the patients is given to the proposed

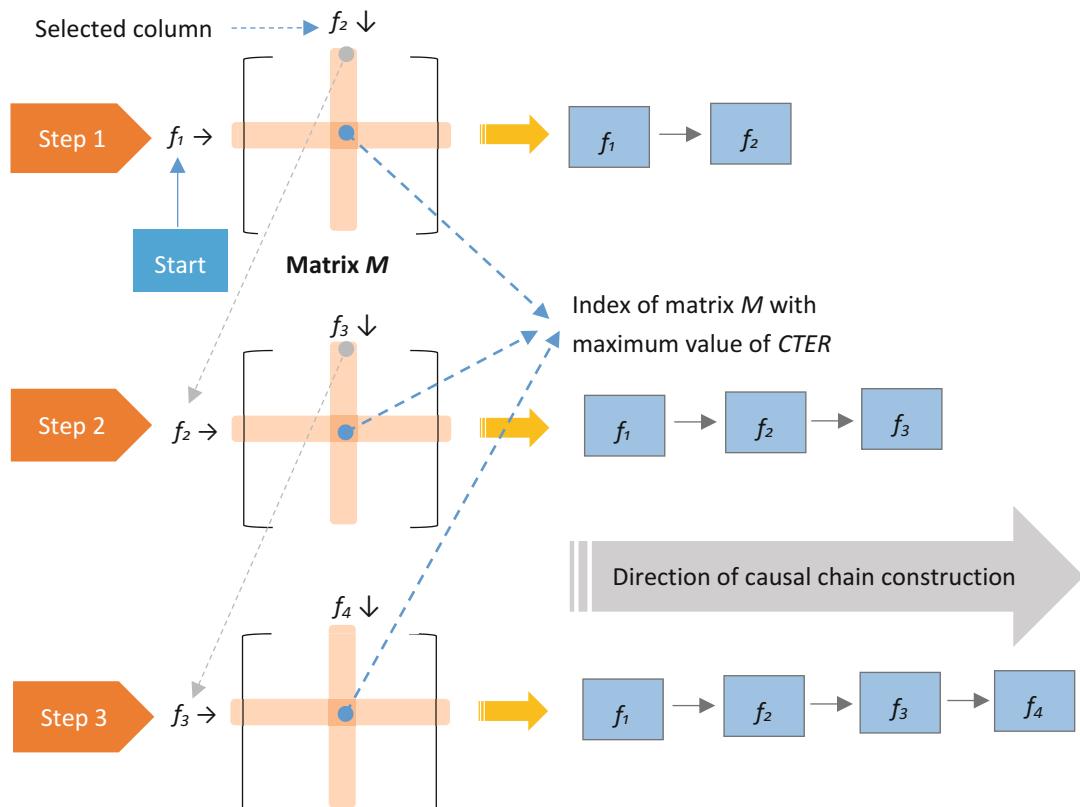


Fig. 2 The process of constructing a causal chain with the length of 4 by row-major approach

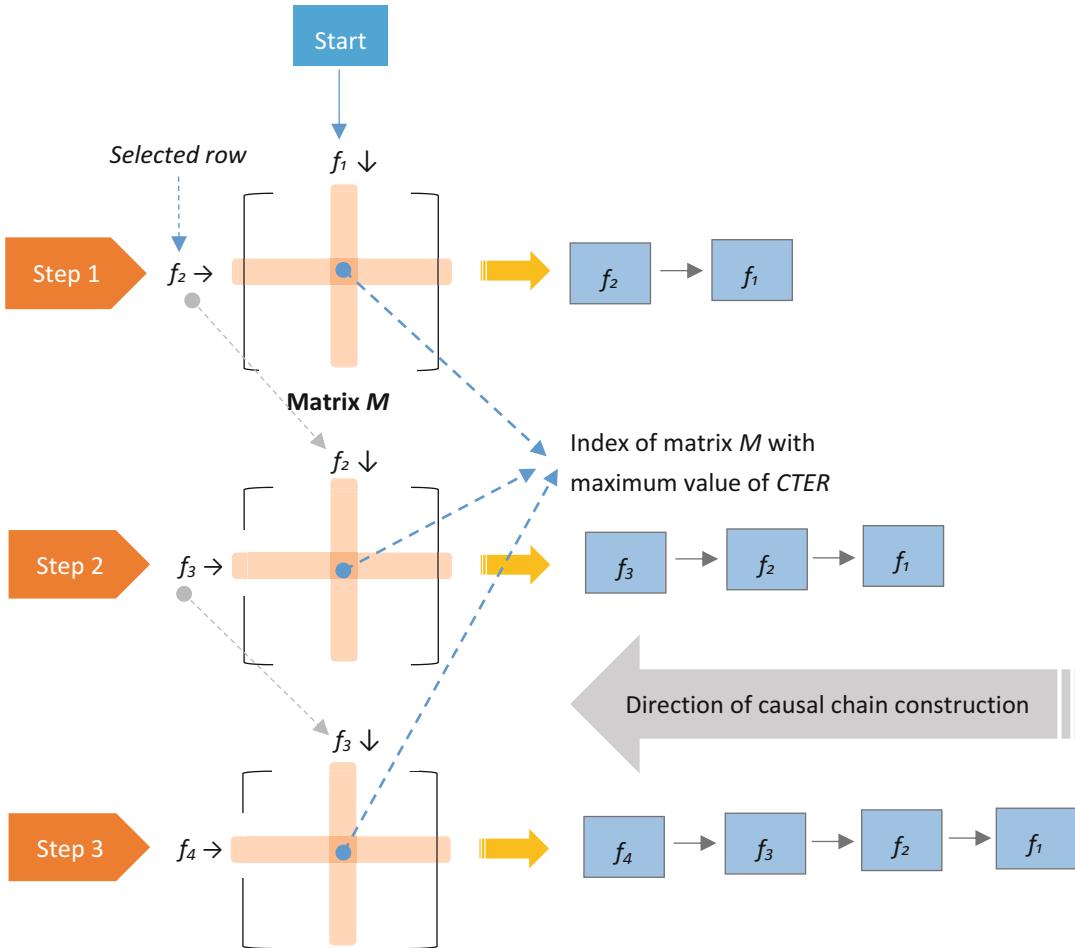


Fig. 3 The process of constructing a causal chain with the length of 4 by column-major approach

method, and after constructing the probabilistic causal graph G , the set of most probable causal chains Π from graph G is extracted. In the PRO-ACT dataset, almost 6000 cases among all 10,000 individuals have non-empty clinical trial information of dynamic features. We used 5000 of the cases for training and 1000 of them for testing the model. In this project, the causal tendency threshold δ and entropy threshold ε are set with the values 0.6 and 4, respectively. Since there are 86 dynamic variables in the PRO-ACT dataset, the dimension of CFD matrix M is $[86 \times 86]$.

The causal graph of ALS has 18 nodes and 25 edges which is demonstrated in Fig. 4. Here, we have 79% of dimension reduction in the size of feature space, because only 18 features from all

86 dynamic features are causal features. Moreover, there are $86 \times 86 = 7396$ potential causal edges wherein only 25 causal edges were selected in the causal graph, and hence, 99.7% of the causal edges were excluded in the causal graph.

The characteristics of all causal edges in the causal graph of Fig. 4 are illustrated in Table 3. The proposed algorithm discovered seven causal chains based on the row-major processing, and all of these causal chains were highlighted with thick arcs in the causal graph of Fig. 4. The proposed algorithm could not find any causal chain based on the column-major approach with a minimum length of three features.

After construction of the causal graph, in the test phase, 3 months' clinical information of 1000

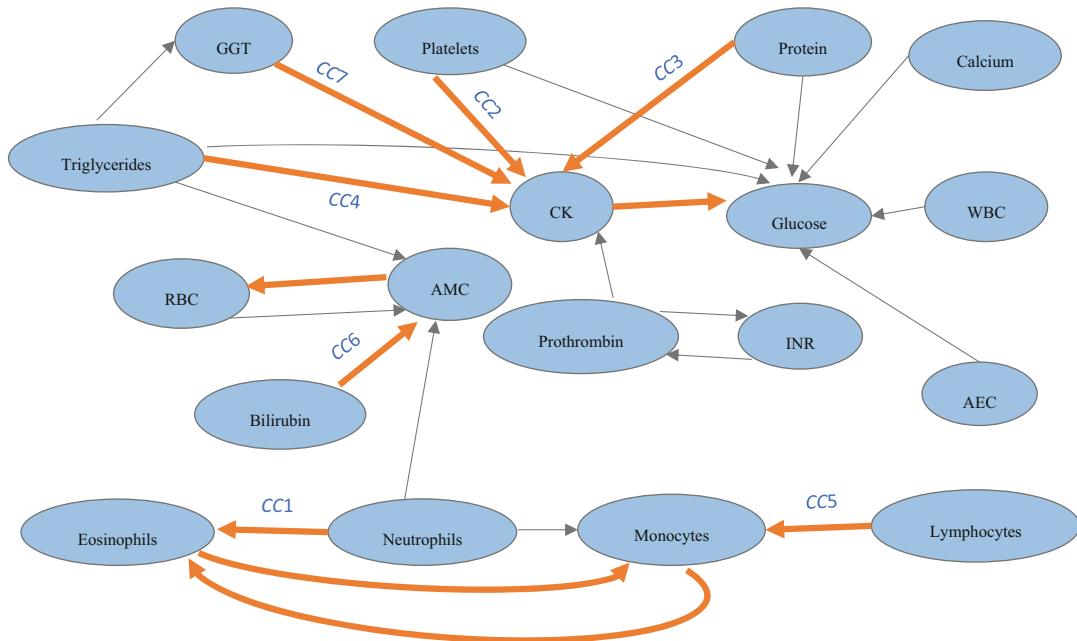


Fig. 4 The causal graph of ALS disease resulted from PCDSD model on the PRO-ACT dataset. The graph includes 18 nodes and 25 edges; most probable causal

chains of the causal graph (CC1 to CC7) are highlighted with orange thick arcs

patients are given to the model, which had not given to the algorithm previously; and the model was expected to predict the rate of ALS progression in the next 9 months (from month 4 to month 12) based on the causal graph G and causal chains Π . Diagram of the average error of ALSFRS prediction in the test patients on the basis of causal graph and causal chains is demonstrated in Fig. 5.

In order to compare the results of the PCDSD model with machine learning algorithms, the slope of ALSFRS for patients is defined as Eq. 1. Slope of ALS patients indicates the rate of ALSFRS value from the first month (m_{first}) to the last month (m_{last}) of prediction [18]. If the value of *slope* was negative, it means that the value of ALSFRS at the beginning of the trial is larger than its value at the end of the trial, and hence because of the progressive nature of ALS disease, the slope value of the patients is often negative:

The root-mean-square error (RMSE) of slope prediction for the PCDSD model and similar methods is shown in Table 4. The results show that Bayesian tree, random forest, and support vector regression outperformed the PCDSD

model. In addition, the PCDSD model (based on causal graph and causal chains) outperformed linear and multivariate regression methods with lower RMSE values.

Conclusion

The AI approach, especially ML methods, has been increasingly used in ALS causation for building diagnostic and prognostic modeling. Current ALS prognostic models rely on clinical features and laboratory tests which might not be sufficient to estimate progression rate, death risk, or survival in patients; they should account for disease heterogeneity rather than solely on the use of ALSFRS-R score to provide more accurate predictions. Study of ALS disease on the PRO-ACT framework as the largest clinical trial dataset of ALS indicates that longitudinal study of the disease with ML methods plays a pivotal role in identifying ALS causation, treatment, and individualized patient care. Discovering causal relationship between factors related to ALS with a

Table 3 Characteristics of all causal edges in the causal graph of Fig. 4

Causal edge	Causal tendency	Cause change	Effect change	Entropy	ALSFRS change	ALSFRS entropy	Men	Time position	Death	Age	CTER
Protein → CK	0.83	-0.04	-0.61	3.09	-0.97	4.4	0.63	0.3	0.23	55.02	0.27
Protein → glucose	0.62	-0.04	0.44	2.75	-1.2	4.41	0.63	0.4	0.23	55.09	0.23
GGT → CK	0.75	-0.16	-0.19	3.93	-0.89	4.23	0.64	0.2	0.23	55.39	0.19
Bilirubin → AMC	0.6	-0.16	-0.49	3.55	-0.47	4.22	0.64	0.2	0.24	55.45	0.17
WBC → glucose	0.6	-0.51	0.42	3.75	-0.83	4.51	0.63	0.79	0.27	54.51	0.16
Neutrophils → monocytes	0.82	0.11	-2.83	2.64	-0.89	4.49	0.62	0.28	0.22	55.72	0.31
Neutrophils → AMC	0.6	0.11	-0.49	2.84	-0.89	4.72	0.61	0.21	0.26	56.08	0.21
Neutrophils → eosinophils	0.92	0.11	0.02	2.8	-1.1	4.4	0.62	0.25	0.22	55.82	0.33
Lymphocytes → monocytes	0.7	0.19	-2.83	2.93	-0.96	4.5	0.62	0.7	0.23	55.02	0.24
Monocytes → eosinophils	0.6	-2.83	0.02	3.13	-1.21	4.1	0.63	0.19	0.22	55.84	0.19
AMC → RBC	0.66	-0.49	-14.173	2.69	-0.47	4.24	0.63	0.21	0.25	55.73	0.24
Eosinophils → monocytes	0.76	-0.39	-0.27	3.1	-1.05	4.18	0.62	0.19	0.21	55.85	0.25
AEC → glucose	0.62	-0.51	0.44	3.6	-0.97	4.25	0.63	0.29	0.25	55.09	0.17
RBC → AMC	0.66	-14.154	-0.49	2.71	-0.47	4.25	0.63	0.2	0.25	55.78	0.24
Platelets → CK	0.83	-0.24	-0.61	3.03	-0.97	4.2	0.63	0.82	0.24	54.95	0.27
Platelets → glucose	0.61	-0.24	0.44	2.66	-1.04	4.21	0.62	0.82	0.25	55.11	0.23
CK → glucose	0.67	-0.59	0.44	3.38	-0.79	4.16	0.63	0.85	0.23	54.91	0.2
Triglycerides → GGT	0.69	0.25	-0.6	3.44	-1.27	4.25	0.63	0.2	0.23	55.19	0.2
Triglycerides → AMC	0.63	0.25	-0.49	3.17	-0.84	4.24	0.63	0.21	0.25	55.46	0.2
Triglycerides → CK	0.97	0.25	-0.33	3.88	-0.54	4.21	0.63	0.2	0.22	55.2	0.25
Triglycerides → glucose	0.67	0.25	0.44	3.12	-1.05	4.21	0.63	0.2	0.23	55.25	0.21
Calcium → glucose	0.61	-0.15	0.44	3.12	-1.22	4.15	0.62	0.35	0.23	55.38	0.2
Prothrombin → CK	0.99	0.03	-0.56	3.92	-0.71	4.56	0.64	0.92	0.15	52.44	0.25
Prothrombin → INR	1.17	0.02	-0.11	3.95	-0.77	5.37	0.64	0.88	0.02	54.13	0.3
INR → prothrombin	1.61	-0.07	-0.14	3.87	-0.82	5.37	0.65	0.93	0.02	54.13	0.42

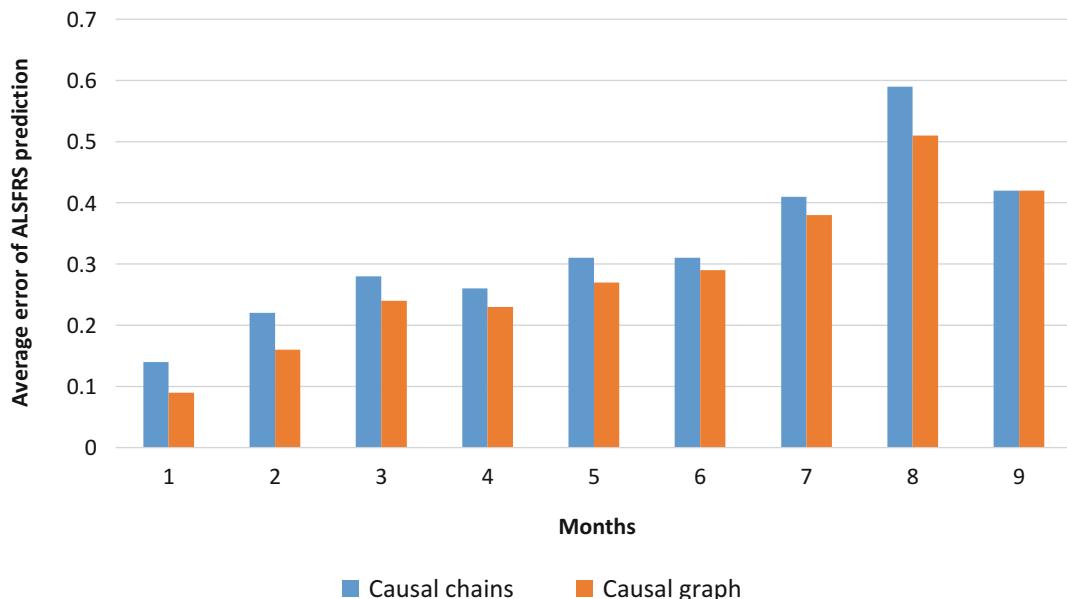


Fig. 5 Diagram of the average error of ALSFRS prediction based on causal graph and causal chains resulted from PCDSD model

Table 4 RMSE values for predicting ALSFRS slope. Comparison between PCDSD model and some machine learning algorithms in the crowdsourcing competition [18]. Bayesian tree and linear regression had minimum and maximum error values in this prediction

Methods	RMSE
Bayesian trees	0.544
Random forest	0.559
Support vector regression	0.559
PCDSD model (causal graph)	0.854
PCDSD model (causal chains)	0.882
Multivariate regression	0.89
Linear regression	1.30

probabilistic approach could bring us closer to the origin of the disease. By knowing causal relations between biological and clinical factors of ALS, neuroscientists would be able to have better understanding of the disease in order to reduce the disease progression rate. Population-based epidemiological studies of ALS show that incidence of patients varies in different countries so that the incidence of ALS in countries with different ancestral population is lower than countries with homogenous population [2, 31, 32]. Since PRO-ACT is US-based dataset of ALS, based on

heterogeneity and epidemiological analyses of ALS disease, the findings from the PRO-ACT dataset could not be generalized to whole ALS population in the world. Thus, a comprehensive dataset with admixed population all over the world is required to fully understand the ALS evolution.

Acknowledgments This work was supported by the Italian Ministry of Health (Ministero della Salute, Ricerca Sanitaria Finalizzata, grant RF-2016-02362405); the Progetti di Rilevante Interesse Nazionale program of the Ministry of Education, University, and Research (grant 2017SNW5MB); the European Commission's Health Seventh Framework Programme (FP7/2007–2013 under grant agreement 259867); and the Joint Programme-Neurodegenerative Disease Research (Strength, ALS-Care, and Brain-Mend projects), granted by Italian Ministry of Education, University and Research. This study was performed under the Department of Excellence grant of the Italian Ministry of Education, University, and Research to the “Rita Levi Montalcini” Department of Neuroscience, University of Torino, Italy.

Declaration of Competing Interests Adriano Chiò serves on the Scientific Advisory Board for Mitsubishi Tanabe, Roche, Biogen, Denali, and Cytokinetics.

References

1. Hardiman O, Al-Chalabi A, Chio A, Corr EM, Logroscino G, Robberecht W, et al. Amyotrophic lateral sclerosis. *Nat Rev Dis Primers*. Nature Publishing Group. 2017;3:1–19.
2. Chiò A, Logroscino G, Traynor BJ, Collins J, Simeone JC, Goldstein LA, et al. Global epidemiology of amyotrophic lateral sclerosis: a systematic review of the published literature. *Neuroepidemiology*. Karger Publishers. 2013;41:118–30.
3. Grollemund V, Pradat P-F, Querin G, Delbot F, Le Chat G, Pradat-Peyre J-F, et al. Machine learning in amyotrophic lateral sclerosis: achievements, pitfalls, and future directions. *Front Neurosci. Frontiers*. 2019;13:135.
4. Bede P. From qualitative radiological cues to machine learning: MRI-based diagnosis in neurodegeneration. *Future Med*. 2017;5:8.
5. Cedarbaum JM, Stambler N, Malta E, Fuller C, Hilt D, Thurmond B, et al. The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. *J Neurol Sci*. Elsevier. 1999;169:13–21.
6. Al-Chalabi A, Hardiman O. The epidemiology of ALS: a conspiracy of genes, environment and time. *Nat Rev Neurol*. Nature Publishing Group. 2013;9:617.
7. Chiò A, Moglia C, Canosa A, Manera U, D’Ovidio F, Vasta R, et al. ALS phenotype is influenced by age, sex, and genetics: a population-based study. *Neurology*. AAN Enterprises. 2020;94:1–9.
8. Turner MR, Hardiman O, Benatar M, Brooks BR, Chio A, De Carvalho M, et al. Controversies and priorities in amyotrophic lateral sclerosis. *Lancet Neurol*. Elsevier. 2013;12:310–22.
9. Schuster C, Hardiman O, Bede P. Survival prediction in Amyotrophic lateral sclerosis based on MRI measures and clinical characteristics. *BMC Neurol*. BioMed Central. 2017;17:1–10.
10. van der Burgh HK, Schmidt R, Westeneng H-J, de Reus MA, van den Berg LH, van den Heuvel MP. Deep learning predictions of survival based on MRI in amyotrophic lateral sclerosis. *NeuroImage: Clin*. Elsevier. 2017;13:361–9.
11. Atassi N, Berry J, Shui A, Zach N, Sherman A, Sinani E, et al. The PRO-ACT database design, initial analyses, and predictive features. *Neurology*. 2014;83:1719–25.
12. Seibold H, Zeileis A, Hothorn T. Individual treatment effect prediction for amyotrophic lateral sclerosis patients. *Stat Methods Med Res*. SAGE Publications Sage UK: London, England. 2018;27:3104–25.
13. Ong M-L, Tan PF, Holbrook JD. Predicting functional decline and survival in amyotrophic lateral sclerosis. *PLoS One*. Public Library of Science San Francisco, CA USA. 2017;12:e0174925.
14. Jahandideh S, Taylor AA, Beaulieu D, Keymer M, Meng L, Bian A, et al. Longitudinal modeling to predict vital capacity in amyotrophic lateral sclerosis. *Amyotroph Lateral Scler Frontotemporal Degener*. Taylor & Francis. 2018;19:294–302.
15. Huang Z, Zhang H, Boss J, Goutman SA, Mukherjee B, Dinov ID, et al. Complete hazard ranking to analyze right-censored data: an ALS survival study. *PLoS Comput Biol*. Public Library of Science. 2017;13:e1005887.
16. Ahangaran M, Jahed-Motlagh MR, Minaei-Bidgoli B. Causal discovery from sequential data in ALS disease based on entropy criteria. *J Biomed Inform*. 2019;89:41–55.
17. Ahangaran M, Jahed-Motlagh MR, Minaei-Bidgoli B. A novel method for predicting the progression rate of ALS disease based on automatic generation of probabilistic causal chains. *Artif Intell Med*. Elsevier. 2020;107:101879.
18. Küffner R, Zach N, Norel R, Hawe J, Schoenfeld D, Wang L, et al. Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. *Nat Biotechnol*. 2015;33:51–7.
19. Renton AE, Chiò A, Traynor BJ. State of play in amyotrophic lateral sclerosis genetics. *Nat Neurosci*. 2014;17:17.
20. Cedarbaum JM, Stambler N. Performance of the amyotrophic lateral sclerosis functional rating scale (ALSFRS) in multicenter clinical trials. *J Neurol Sci*. 1997;152:s1–9.
21. Wijesekera LC, Leigh PN. Amyotrophic lateral sclerosis. *Orphanet J Rare Dis*. Springer. 2009;4:3.
22. Cutter GR, Baier ML, Rudick RA, Cookfair DL, Fischer JS, Petkau J, et al. Development of a multiple sclerosis functional composite as a clinical trial outcome measure. *Brain*. Oxford University Press. 1999;122:871–82.
23. Romero K, De Mars M, Frank D, Anthony M, Neville J, Kirby L, et al. The coalition against major diseases: developing tools for an integrated drug development process for Alzheimer’s and Parkinson’s diseases. *Clin Pharmacol Ther*. Wiley Online Library. 2009;86:365–7.
24. PRO-ACT data set [Internet]. www.ALSDatabase.org. Accessed Sept 2020.
25. Bishop CM. *Pattern recognition and machine learning*. Springer; 2006.
26. Schuster C, Hardiman O, Bede P. Development of an automated MRI-based diagnostic protocol for amyotrophic lateral sclerosis using disease-specific pathognomonic features: a quantitative disease-state classification study. *PLoS One*. Public Library of Science San Francisco, CA USA. 2016;11:e0167331.
27. Querin G, El Mendili M-M, Bede P, Delphine S, Lenglet T, Marchand-Pauvert V, et al. Multimodal spinal cord MRI offers accurate diagnostic classification in ALS. *J Neurol Neurosurg Psychiatry*. BMJ Publishing Group Ltd. 2018;89:1220–1.
28. Taylor AA, Fournier C, Polak M, Wang L, Zach N, Keymer M, et al. Predicting disease progression in

- amyotrophic lateral sclerosis. *Ann Clin Transl Neurol*. Wiley Online Library. 2016;3:866–75.
29. Hothorn T, Jung HH. RandomForest4Life: a random forest for predicting ALS disease progression. *Amyotroph Lateral Scler Frontotemporal Degener*. Taylor & Francis. 2014;15:444–52.
30. Shannon CE. A mathematical theory of communication. *Bell Syst Techn J* [Internet]. 1948;27:379–423. <http://cm.bell-labs.com/cm/ms/what/shannonday/shanon1948.pdf>
31. Logroscino G, Traynor BJ, Hardiman O, Chiò A, Mitchell D, Swingler RJ, et al. Incidence of amyotrophic lateral sclerosis in Europe. *J Neurol Neurosurg Psychiatry*. BMJ Publishing Group Ltd. 2010;81:385–90.
32. Gordon PH, Mehal JM, Holman RC, Rowland LP, Rowland AS, Cheek JE. Incidence of amyotrophic lateral sclerosis among American Indians and Alaska natives. *JAMA Neurol*. American Medical Association. 2013;70:476–80.



AIM in Ménière's Disease

122

Young Sang Cho and Won-Ho Chung

Contents

Introduction	1706
Overview of Ménière's Disease	1706
The History of Inner Ear MRI to Visualize Endolymphatic Space and Hydrops	1708
Current Diagnostic Method and Dilemma in Ménière's Disease	1708
Sequence and Analysis of Inner Ear MRI for the Diagnosis of Ménière's Disease	1708
Development of Artificial Intelligence in the Medical Field: Focusing on Medical Image Analysis	1709
The Use of Artificial Intelligence in Ménière's Disease	1710
The Future of Artificial Intelligence in Ménière's Disease	1714
References	1715

Abstract

Ménière's disease (MD) is difficult to diagnose objectively and evaluate the treatment outcomes. Although pure tone audiometry is the only objective test included in the diagnostic criteria, inner ear MRI technique, which was recently developed to visualize endolymphatic hydrops (EH), is useful for the diagnosis of MD. However, analyzing methods are reported

to be diverse, and sometimes, they are time-consuming and complicated. In recent years, the rapidly developing field of artificial intelligence (AI) showed outstanding performance in image recognition. In particular, convolutional neural network (CNN) based on deep learning plays a remarkable role in today's medical field, where imaging analysis is critical. We developed a CNN-based deep learning model called INHEARIT (INner ear Hydrops Estimation via ARtificial InTelligence) for automatic calculation of EH ratio in a segmented region of the cochlea and vestibule. The model can generate results that are highly consistent with those generated by manual calculation more

Y. S. Cho · W.-H. Chung (✉)
Department of Otorhinolaryngology-Head and Neck
Surgery, Samsung Medical Center, Sungkyunkwan
University School of Medicine, Seoul, South Korea

quickly. This automated analysis of inner ear MRI using deep learning would be useful for diagnosis and follow-up of MD. It is also expected to be widely used in differential diagnosis of various EH-related diseases.

Keywords

Ménière's disease · Endolymphatic hydrops · MRI · Deep learning · Hearing loss · Vertigo · Dizziness · Artificial intelligence · Convolutional neural network

Introduction

Ménière's disease (MD) is a multifactorial disorder with typical symptoms of recurrent attacks of vertigo, fluctuating hearing loss, tinnitus, and sensations of ear fullness. Endolymphatic hydrops (EH) is a histologic hallmark of MD in which the endolymphatic spaces in the cochlea and vestibule are distended [1]. According to a 1995 consensus statement from the Committee on Hearing and Equilibrium of the American Academy of Otolaryngology-Head and Neck Surgery (AAO-HNS), “certain” MD can be confirmed only by the histological demonstration of EH in postmortem temporal bone specimens [2]. Therefore in 2015, a committee of the Bárány Society revised the diagnostic criteria to remove the concept of “certain MD” [3] (Table 1).

According to the diagnostic criteria, pure tone audiometry (PTA) is the only objective test used to diagnose a “definite” or “probable” MD. Other measures, such as electrocochleography (EcoG) and vestibular evoked myogenic potential (VEMP), have been used as adjunctive diagnostic tools for more than 30 years [4] as it is impossible to visualize EH but only estimate [5]. However, it is occasionally challenging to diagnose MD using only patients' symptoms.

With the advancement of imaging technology, EH can be visualized by MRI as an objective marker in patients with MD [6]. Many studies have been published regarding EH visualized by intratympanic or intravenous injection of a contrast media in MRI. However, these studies used

various methods for analysis, and sometimes, quantitative analysis takes considerable time and effort to calculate the correct hydrops ratio.

In recent years, with the remarkable development of deep learning, artificial intelligence (AI), including machine learning, is widely used in various fields of medical science [7]. Among them, image analysis using a convolutional neural network is rapidly growing [8]. This can be of great help for an automatic evaluation of EH in patients with MD using inner ear MRI. This chapter will describe the role of AI in inner ear MRI which was recently applied for diagnosing MD.

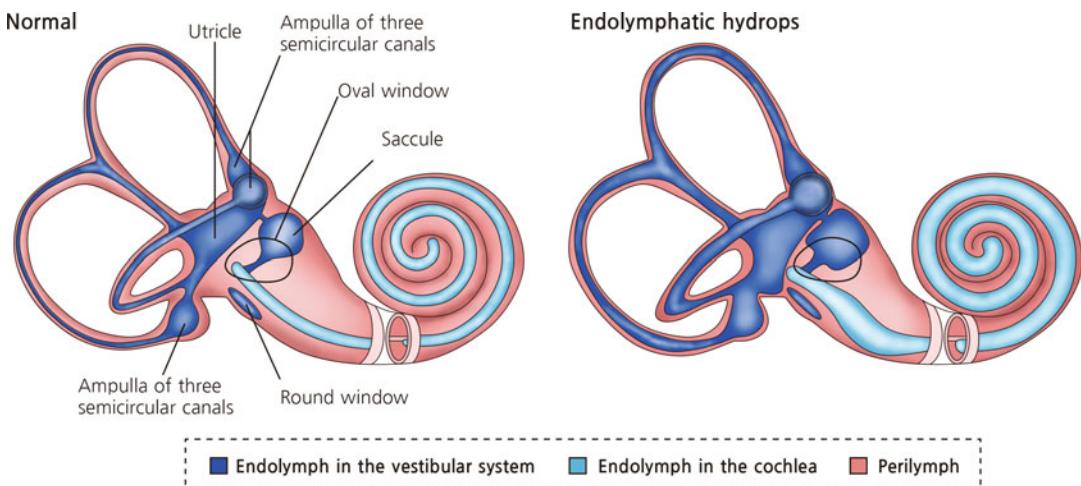
Overview of Ménière's Disease

Ménière's disease (MD) is named after a French doctor, Prosper Ménière (1799–1862), in the middle of the nineteenth century [9]. Recurrent dizziness and fluctuating hearing loss, tinnitus, and ear fullness are typical symptoms. He claimed MD as an inner ear disorder causing dizziness through an autopsy of a patient who complained severe recurrent dizziness during his lifetime. However, he did not provide any clear pathophysiology of MD [10]. It was not until 1938, 70 years later, that the Yamakawa group and the Hallpike and Cairns group almost simultaneously presented the pathophysiological evidence of MD. They found endolymphatic hydrops (EH) in human temporal bone specimens with dilated scala media in the cochlea, which was later considered a histopathological marker of MD (Fig. 1) [1, 11].

The epidemiologic studies on the prevalence of MD vary. This is thought to be due to the uncertainty in the diagnostic criteria and pathophysiological mechanism. The prevalence rate is reported to vary from 17 to 513 per 100,000 people [12]. MD is a complex and heterogeneous disease in which various factors may interact, including genetics, immunity, hydrodynamics in the inner ear, psychological aspect, and cellular and molecular mechanisms [6]. The causes or course of the disease can be diverse, but one of the most important

Table 1 AAO-HNS 1995 criteria and 2015 proposed new criteria for Ménière's disease

	AAO-HNS 1995 criteria	2015 new criteria by Barany Society
Certain Ménière's disease	Definite Ménière's disease, plus histopathological confirmation	Deleted
Definite Ménière's disease	At least two definite spontaneous episodes of vertigo lasting at least 20 min Audiometrically documented hearing loss on at least one occasion Tinnitus or aural fullness in the treated ear Other causes excluded	At least two spontaneous episodes of vertigo, each lasting from 20 min to 12 h Audiometrically documented low-frequency to medium-frequency sensorineural hearing loss in one ear, defining the affected ear on at least one occasion before, during, or after one of the episodes of vertigo Fluctuating aural symptoms (hearing, tinnitus, or fullness) in the affected ear Not better accounted for by another vestibular diagnosis
Probable Ménière's disease	One definitive episode of vertigo Audiometrically documented hearing loss on at least one occasion Tinnitus or aural fullness in the treated ear Other causes excluded	At least two episodes of vertigo or dizziness, each lasting from 20 min to 24 h Fluctuating aural symptoms (hearing, tinnitus or fullness) in the affected ear Not better accounted for by another vestibular diagnosis
Possible Ménière's disease	Episodic vertigo of the Ménière type‡ without documented hearing loss Sensorineural hearing loss, fluctuating or fixed, with Other causes excluded disequilibrium, but without def	Deleted

**Fig. 1** Endolymphatic hydrops (EH), a histologic hallmark observed in Ménière's disease. EH increases the endolymphatic space, which can affect the inner ear and cause symptoms. Progression of MD is associated with progression of EH

histopathological hallmarks is the distention of the endolymphatic space in the inner ear. For treatment, lifestyle modifications, such as stress reduction, dietary restrictions, and medical treatments, are mainly used. Surgical treatments including vestibular ablation or sac

decompression are available in medically intractable cases. Patients may experience psychiatric problems along with other symptoms as poor quality of life is expected due to unexpected vertigo attacks and severe hearing loss [13].

The History of Inner Ear MRI to Visualize Endolymphatic Space and Hydrops

In the twentieth century, with the development of imaging technology using MRI, various efforts have been attempted to identify EHendolymphatic hydrops (EH) in vivo. In 2004, Duan et al. succeeded in visualizing EH in vivo for the first time in a guinea pig using 4.7 T MRI [14]. Nakashima et al. succeeded in confirming EH after injecting contrast media through intratympanic (IT) and intravenous (IV) injections into MD patients using 3 T MRI [15]. Recently, many reports have been published regarding the use of MRI to assess EH. In particular, IV gadolinium (Gd)-enhanced inner ear MRI has shown promising results [16, 17]. IV-Gd inner ear MRI is less invasive and is much more efficient because it requires less time after the contrast agent injection compared to the IT method (4 h vs. 24 h) and can evaluate both sides simultaneously [18]. Direct assessment of EH using MRI can be applied in clinical practice using the semiquantification and grading protocol suggested by Naganawa et al., which is currently the most widely used [16]. Several studies have shown a relationship between the severities of EH in MRI and MD patients' symptom correlation, and IV-Gd inner ear MRI is beneficial for diagnosing MD by demonstrating the correlation of hydrops with audio-vestibular results [19]. MD research using MRI is still ongoing in many groups, and although there are various methods in visualizing EH, the most widely used sequence is 3D-FLAIR, which uses an intravenous infusion of a gadolinium-based contrast agent. Recently, several different studies evaluate hydrops either by precisely dividing each area of the cochlea and vestibule or by implementing a 3D model to measure the whole volume of the entire endolymphatic space through MRI [19, 20].

Current Diagnostic Method and Dilemma in Ménière's Disease

The diagnostic criteria for MD proposed by AAO-HNS have been widely used and recently revised in 2015 (Table 1). According to the

diagnostic criteria, the diagnosis is inevitably dependent on the subjective symptoms except for pure tone audiometry, and all other vestibular diseases must be excluded.

Besides pure tone audiometry, electrocochleography (EcoG), vestibular evoked myogenic potential (VEMP) test, caloric test, and video head impulse test (vHIT) test have been used as clinical tests. However, it is challenging to accurately assess EH levels with these tests. Among these tests, EcoG has been used for more than 30 years in the diagnosis of MD [4]. However, EcoG is used only as a reference examination to diagnose MD because it cannot directly show the endolymphatic space. Even the diagnostic value of EcoG is still part of a debate in literature [21]. Regarding cervical VEMP (cVEMP) indicating saccular function, there was no significant correlation between the hydrops level and interaural difference (IAD) ratio. Several studies have shown that cVEMP response was reduced or enhanced in MD patients depending on the duration and stage of MD [22, 23]. Previous literature suggested that enhanced cVEMP response in the affected ear was often associated with the early stage of MD, and as MD progresses, a decrease in amplitude was observed. In patients with MD, canal paresis is sometimes observed in caloric testing. However, even with canal paresis, vHIT often shows normal findings. Suppose the hydroptic expansion of the endolymphatic duct in MD patients allows a local flow within the duct. In that case, this could dissipate the hydrostatic pressure caused by thermally induced density difference and diminish or eliminate the cupula's deflection. Thus, the hydrops of the endolymphatic duct could cause a caloric deficit without compromising the vestibulo-ocular reflex response in vHIT [24]. Therefore, additional tests currently used for MD diagnosis do not reflect the level of EH directly, which is a histologic hallmarker.

Sequence and Analysis of Inner Ear MRI for the Diagnosis of Ménière's Disease

Since the mid-2000s, 3 T MRI using a three-dimensional fluid-attenuated inversion recovery (3D-FLAIR) sequence has been used to detect

EH in MD [15]. These days, 3D-FLAIR and inversion recovery turbo spin echo with real reconstruction (3D real-IR) sequence are used the most as an MRI protocol to identify the signal differences between the perilymph (with contrast) and endolymph (without contrast) [25]. A primary advantage of the 3D real-IR sequence is that it describes the surrounding bone as an intermediate (gray) signal, so perilymph, endolymph, and adjacent bone can be distinguished in a single image [20]. To maximize the signal MR imaging was obtained optimally on a 3 T MRI system with a dedicated head coil and multiple receiver channels (over 32 channels). The basic MRI scan sequence is as follows: (1) heavily T2-weighted (hT2W) MR cisternography (MRC) for anatomical reference of total endolymphatic and perilymphatic fluid, (2) hT2W-3D-FLAIR with inversion time of 2250 ms (positive perilymph image (PPI)) to visualize the perilymph space, and (3) hT2W-3D-IR with inversion time of 2050 ms (positive endolymph image (PEI)) for evaluating the endolymphatic space. The parameters of PEI were the same as that of PPI except that the inversion time was 2050 ms. To facilitate comparison, MR cisternography, PPI, and PEI employed identical FOV, matrix size, and slice thickness. The next step is to subtract PEI from the PPI to create HYDROPS (HYbriD of Reversed image Of Positive endolymph signal and native image of positive perilymph Signal) on the scanner console. Upon completion, it is possible to analyze the extent of the endolymphatic space expansion visually. Clinically, several grading systems have been introduced for reading and diagnosis, but the most widely used grading system was announced in 2009 by T. Nakashima et al. (Table 2) [26]. For an accurate quantitative analysis, additional computer software must be used, and this takes a lot of time and effort.

In order to calculate the ratio, a technique for increasing the contrast is also used. To increase the contrast-to-noise ratio of HYDROPS images, HYDROPS-Mi2 images were generated on a DICOM viewer by multiplication of HYDROPS and MRC images (Fig. 2). Additionally, all cochlear and vestibule boundaries must be drawn manually along the contour on MRI by physicians (Fig. 3). This manual process is very

Table 2 Grading system of endolymphatic hydrops using MRI [26]

Grade of hydrops	Vestibule (area ratio ^a)	Cochlea
None	$\leq 33.3\%$	No displacement of Reissner's membrane
Mild	$> 33.3\%, \leq 50\%$	Displacement of Reissner's membrane area of cochlear duct \leq area of the scala vestibuli
Significant	$> 50\%$	Area of the cochlear duct exceeds the area of the scala vestibuli

^aRatio of the area of the endolymphatic space to that of the fluid space (sum of the endolymphatic and perilymphatic spaces) in the vestibule measured on tracings of images

time-consuming, is cumbersome, and is inefficient for clinical settings. An automated analysis system could be an excellent option to accurately calculate EH ratios in real time without a time-consuming and complicated process.

Development of Artificial Intelligence in the Medical Field: Focusing on Medical Image Analysis

The concept of AI was first introduced in the 1950s, but with its limitations in algorithms, data, and computing resources, it was not until the 2000s when AI re-emerged and gained popularity. In particular, the gradient vanishing problem of backpropagation was solved with a new activation function, and deep neural networks based on big data and computing resources, including a graphics processing unit (GPU), showed excellent performance. Along with this, a lot of research and social interest sparked a period of the revival of AI.

Particularly, radiology is the most important in the medical AI field [27]. This is due to developing a convolutional neural network (CNN), which shows excellent image recognition among various deep neural network algorithms. CNN is a deep neural network that mimics the human visual perception process, originating from the neo-cognition model first proposed by Kunihiko Fukushima in 1979 [28]. CNN supports convolution, which has the advantage of being able to

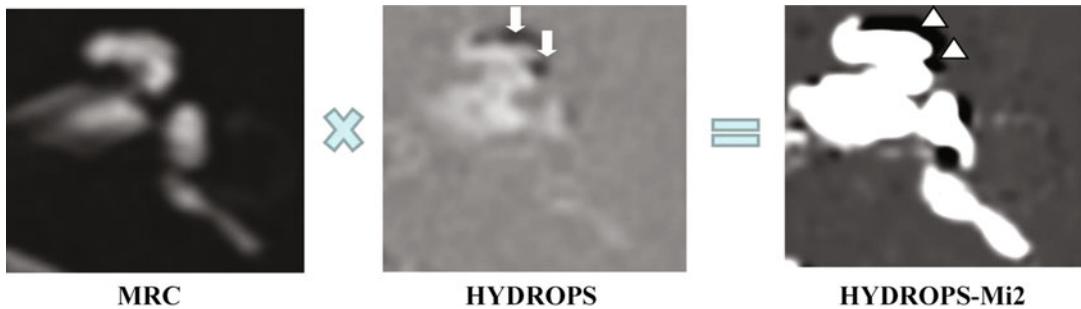


Fig. 2 By multiplying the MRC and the hydrops image, a HYDROPS-Mi2 image with maximized contrast can be created. The endolymphatic space (white arrow) observed

train high-dimensional data with relatively few parameters. This makes it easier to train more than two-dimensional data than general neural networks. Typically, in a conventional artificial neural network (ANN), each neuron in a layer connects to all neurons in the next layer, and each connection is a network parameter [29]. This can result in the creation of a huge number of parameters. Instead of using fully connected layers, CNN is composed of a convolution layer and a pooling layer that extract features of high-dimensional data and a fully connected layer that finally classifies the data. The convolution and pooling layers have the advantage that the number and order of each layer can be arbitrarily adjusted according to the problem the user wants to solve (Fig. 4).

This CNN showed its remarkable progress through the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), which started in 2010. In 2012, Professor Jeffrey Hint's team created an “AlexNet” consisting of 7 hidden layers, 650,000 neurons, 60 million variables, and 630 million network connections at this competition, lowering the error rate to 16%. In 2015, ResNet, which reduced the error rate to 3.57%, appeared, proving that AI outperformed humans (error rate 5%) in the field of image recognition.

Today, these deep neural network models can be used for medical image classification, segmentation, object detection, registration, and other tasks [30]. Chest X-ray, mammography, and computed tomography (CT) in the head area are mainly studied and used most actively [8]. MRI

in the hydrops image is more visible in the Mi2 image (white arrowhead)

is a powerful and noninvasive tool that produces high-quality 3D images of complex brain structures [31]. In addition to the development of algorithms to classify Alzheimer’s or autism, quantification of various brain lesions is also actively studied. Patch-wise CNN architecture is commonly used due to the nature of the complex brain region. This is a simple approach to training a CNN algorithm for segmentation. NxN patches around each pixel are extracted from a given image, and the model is trained on these patches and class-labeled to accurately identify classes, such as normal brain and tumor [32]. Recently, the development of 3D images reconstructed based on MR images or 4D imaging using functional MRI is also accelerating. A CNN architecture with patch-based analysis might be beneficial for the field of medicine when there is a high demand of tools that extract information from large datasets.

The Use of Artificial Intelligence in Ménière’s Disease

Nowadays, many reports have been published regarding the use of MRI to assess EH. To be more specific, IV-Gd-enhanced inner ear MRI has shown good results. The degree of EH correlates well with the severity of symptoms and several audio-vestibular test results that have been used before [33, 34]. Although vestibular migraines and MD are not easily differentiated due to overlapping symptoms, inner ear MRI can

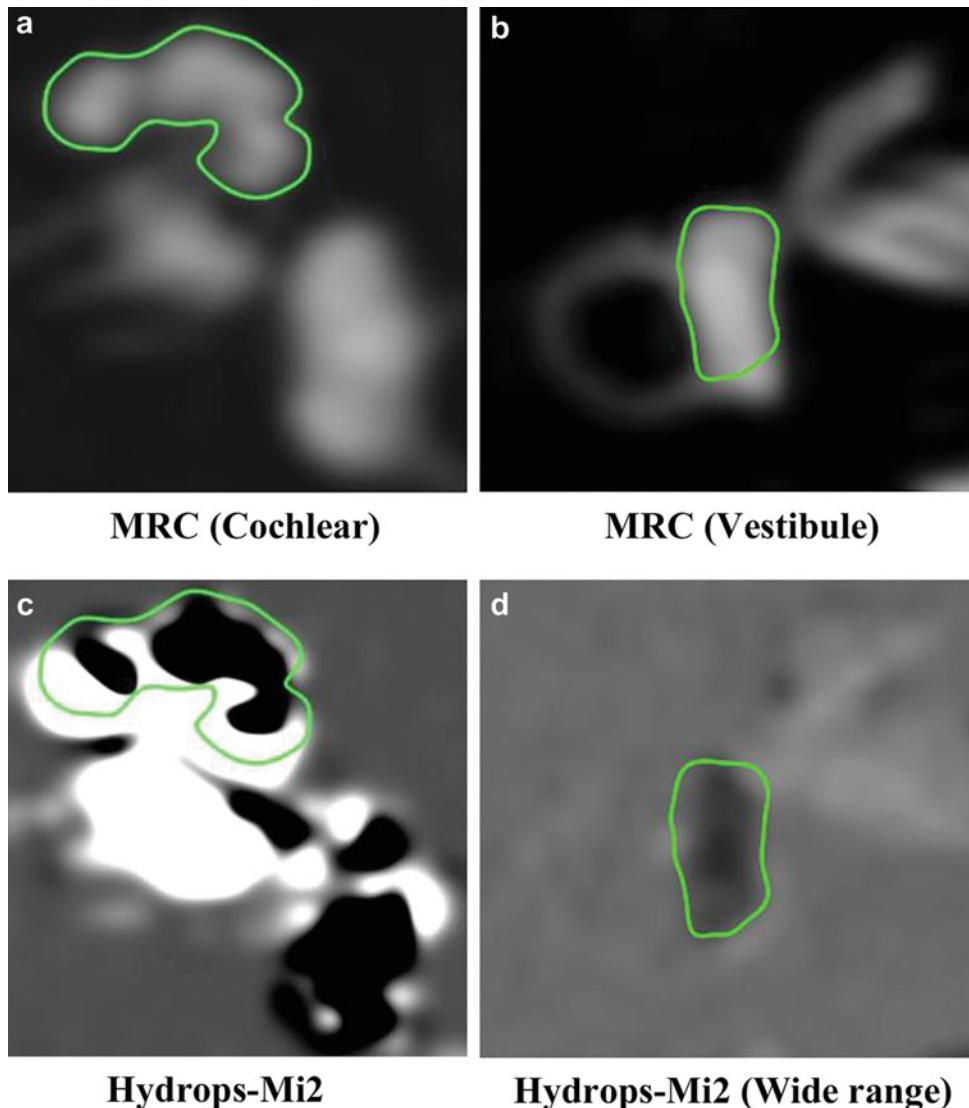


Fig. 3 The process of setting the region of interest (ROI) to calculate the hydrops ratio. **(a)** In the MRC image, the ROI for the cochlear hydrops was drawn along the contour by selecting the slice that best showed the cochlear modiolus. **(b)** For the vestibular ROI, the lowest slice, which the lateral semicircular canal was visualized more than 240°, was selected, and the ampulla was excluded when drawing ROI on MRC. **(c)** A copy of the cochlear

ROI from the MRC image was pasted to the HYDROPS-Mi2 image, and the hydrops ratio was calculated by dividing the number of pixels having negative pixels from the whole pixels. **(d)** A copy of the vestibular ROI from the MRC image was pasted in the HYDROPS-Mi2 (wide range) image. Vestibular hydrops with a negative signal inside is clearly observed

clearly distinguish between the two [35]. Many findings have differentiated inner ear diseases associated with EH (hydropic inner ear disease) using inner ear MRI. For this purpose, it is necessary to accurately and consistently calculate the EH ratio. However, as manual calculations are

required, it is time-consuming to calculate the exact EH ratio.

HYDROPS or HYDROPS-Mi2 images should be created using a specific brand of image viewer software. In addition, all cochlear and vestibule boundaries must be drawn manually along the

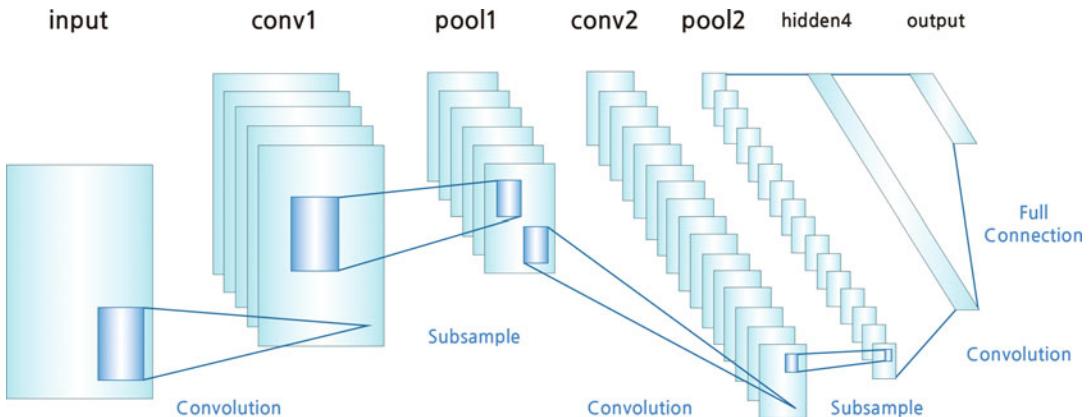


Fig. 4 The architecture of a convolutional neural network

contour on MRI. An automated analysis system could be a good option to accurately calculate the EH ratios in real time. Previous studies have evaluated the automatic segmentation of inner ear organs. For example, one study described semi-automatic CT image segmentation, which combines thresholding techniques and manual segmentation, but experts still needed to localize some points for the segmentation process [36]. Other groups applied a random forest classifier and a Niblack segmentation algorithm to a 3D reconstructed image and measured the endolymph and total fluid space [37]. There is a need for more automated deep learning algorithms for semantic segmentation of individual organs (cochlea and vestibule).

The simplest approach would be to use a fully connected ANN, but this would be highly computationally expensive because every pixel is linked to every neuron. A CNN solves this issue by filtering the connections by proximity (i.e., each neuron accepts inputs from a subsection relative to the receptive field in the image) of the lower layer, making it computationally manageable [38]. Besides, subsection-based processing mimics how individual cortical neurons function (a small portion of the visual field), where components of the CNN operate on local input regions. Accordingly, CNNs have demonstrated good performance in semantic segmentation in natural images as well as medical images [39].

In order to accurately measure EH with CNN, the following steps are required: (1) data preparation for training, (2) accurate segmentation of inner ear organs (cochlea and vestibule) based on the trained images, and (3) accurate and automatic calculation of the EH ratio for the segmented area (Fig. 5).

For accurate segmentation, several factors must be premised. Firstly, as with all CNN-based image analysis, it is necessary to acquire and learn as many datasets as possible to accurately segment the inner ear organ. Due to the specificity of the disease and the characteristics of the MRI protocol, quantitative limitations of the image dataset are inevitable. However, in MR images, it is possible to amplify the original images into the training dataset using flipping, intensity changing, and random shift cropping. The following is an example of a protocol for augmentation of a dataset lacking in the process: (1) low augmentation by flipping and random shifting (144 times); (2) moderate augmentation by flipping, random shifting, and 10 steps of brightness change (1584 times); and (3) high augmentation by flipping, shifting, and 1 step of brightness change (14,544 times). According to this method, it is possible to make a single MR image into more than 14,000 images and train that images. Secondly, since brain MR imaging is more complex and heterogeneous than other organs, a technique that can accurately segment the target inner ear organs is needed. Here, we can use the commonly

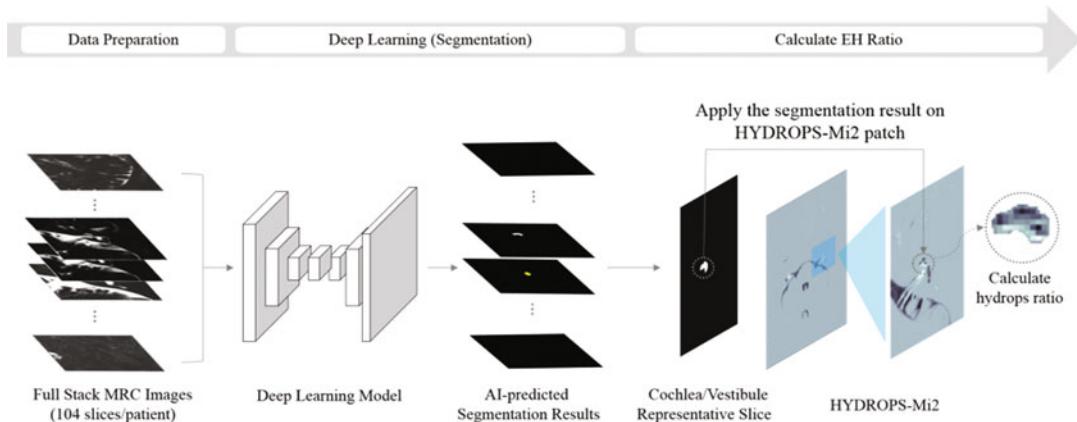


Fig. 5 Process for calculating the endolymphatic hydrops ratio using a convolutional neural network

used learning method of dividing the whole image into small patches. The cochlea and vestibule areas are relatively small according to the original whole MRC images (384×324 pixels). Therefore, the image is cropped 100×100 pixel windows from each side of the inner ear images at the left [215,238] and right [204,92] reference points with the cropping reference points being determined by a radiologist. All the ROIs for the entire dataset resided inside the cropped windows. In addition, the size and location of the inner ear organs are similar regardless of age or sex because they do not grow or change shape after birth,

which is highly beneficial for analysis using CNN [40].

Physicians' annotations on the regions of the cochlea and vestibule using MR cisternography (MRC) images are regarded as the ground truth. The deep learning segmentation model learns from these annotations together with the input MRC images. The results are similar to manual analysis performed by a physician (Fig. 6).

If segmentation is perfectly possible in this way, it is easy to find the area of the endolymphatic space or measure the hydrops ratio. The EH ratio is defined as follows:

$$\text{EH ratio} = \frac{\text{Total number of pixels with negative value in the segmentation area}}{\text{Total number of pixels in the segmentation area}}$$

The negative ratio represents the endolymphatic space area without the perilymph enhanced by a Gd. Recent research shows that the EH ratio is surprisingly consistent between the results measured by the specialist and the results predicted by AI [41]. In this study, the average interclass correlation coefficient (ICC) value for an entire image was 0.971, while the average ICC of the vestibule images (0.980) was higher than the cochlea images (0.952). For accurate calculations, a skilled specialist takes 10 min to calculate the EH ratio through post-processing of images, whereas AI takes only 0.168 s from reading to analysis.

Using this algorithm and framework, we can estimate the EH ratio accurately and quickly. By analyzing hydrops with MRI, the diagnosis of MD could be more accurately made by differentiating MD from other diseases with similar symptoms. In addition, an automatic quantitative analysis of hydrops ratios using inner ear MRI may be applied for assessing the stage of disease and prognosis. Furthermore, it may make the diagnosis of many diseases and symptoms suspected to involve EH much easier. Currently, only the representative section is extracted from the entire MR image, and the EH ratio is calculated. However, it

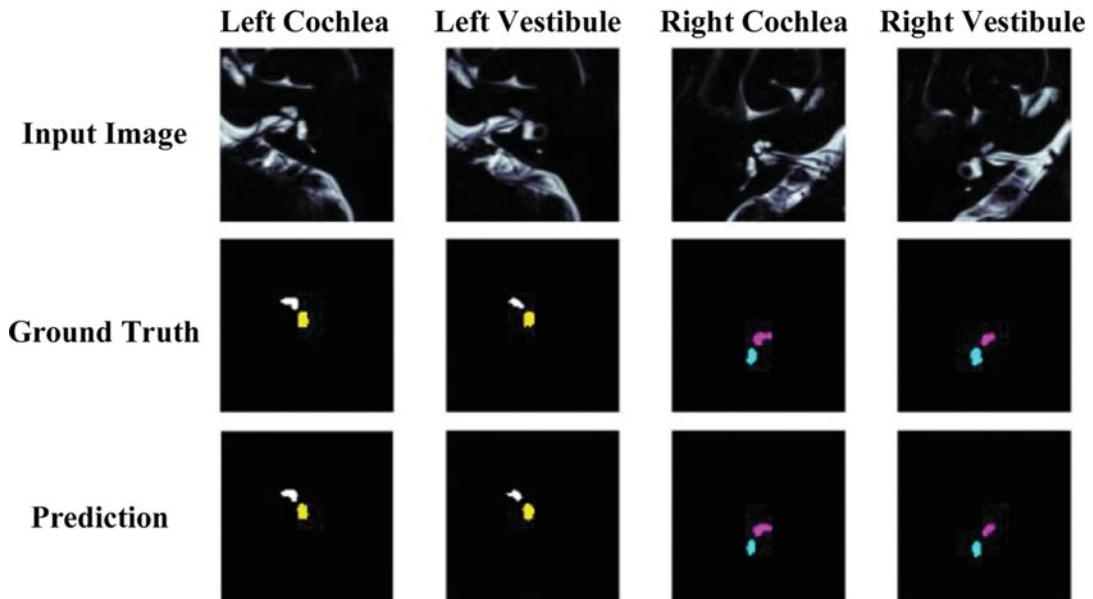


Fig. 6 AI-based segmentation results from a fully annotated dataset. Those examples show that AI-based prediction performs well compared to physicians' annotations (ground truth)

will be more effective if the volume of EH can be predicted quickly in the future.

The Future of Artificial Intelligence in Ménière's Disease

With the development of MRI machines, data and image processing technology, big data, and cloud systems, AI's role in the medical field is bound to increase as time goes by. Until now, algorithms have begun to help clinicians or predict clinical outcomes that are useful in medical systems, but patient-centered algorithms are still lacking [8]. While these technologies can be helpful for diagnosis and treatment of inexperienced physicians, they can lower the barrier to medical knowledge at the same time, making it much easier for patients to have a diagnostic approach to their disease. AI will be used in most of our lives in the future, but we need to be more careful in the medical field. The disease progression of all patients is not always the same, and the fatal errors of generalized algorithms can cause problems that could be more serious than those caused between a single patient and doctors.

Nevertheless, just as evidence-based medicine forms the basis for all medical fields, AI will be an excellent tool to quickly and easily utilize this evidence in the future. Therefore, we can refer to the evidence-based content proposed by AI, but we should still be careful when it comes to utilizing this "evidence-based content" in clinical practice.

Currently, several studies are underway, but the possibility of using inner ear MRI in MD is not yet verified and validated. MRI can confirm the endolymph space and calculate the EH ratio, but the protocol or analysis method differs between researchers. In some studies, symptoms improved through treatment in MD patients, but the pattern of hydrops did not change or rather increased in MRI performed after treatment. Hence, it is necessary to further consider how EH in MD causes symptoms [42, 43].

AI will become important in most medical fields in the future. It seems strongly possible that CNN-based deep learning will be widely used for analysis and reading of medical images. There are still many aspects that need to be improved when it comes to the diagnosis of MD, but AI will be able to perform clustering and

classification by synthesizing patients' symptoms, audio-vestibular tests, and MRI analysis. Healthcare professionals will need to carefully consider and strive for symbiosis between AI and the field of medicine in the future.

References

- Hallpike CS, Cairns H. Observations on the pathology of ménieré's syndrome (section of otology). *Proc R Soc Med.* 1938;31(11):1317–36.
- Committee on Hearing and Equilibrium. Committee on Hearing and Equilibrium guidelines for the diagnosis and evaluation of therapy in Meniere's disease. American Academy of Otolaryngology-Head and Neck Foundation, Inc. *Otolaryngol Head Neck Surg.* 1995;113(3):181–5. [https://doi.org/10.1016/S0194-5998\(95\)70102-8](https://doi.org/10.1016/S0194-5998(95)70102-8).
- Lopez-Escamez JA, Carey J, Chung WH, Goebel JA, Magnusson M, Mandala M, et al. Diagnostic criteria for Meniere's disease. *J Vestibul Res-Equil.* 2015;25(1):1–7. <https://doi.org/10.3233/VES-150549>.
- Claes GM, De Valck CF, Van de Heyning P, Wuyts FL. The Meniere's Disease Index: an objective correlate of Meniere's disease, based on audiometric and electrocochleographic data. *Otol Neurotol.* 2011;32(5):887–92. <https://doi.org/10.1097/MAO.0b013e318219ff9a>.
- Durrant JD, Wang J, Ding DL, Salvi RJ. Are inner or outer hair cells the source of summing potentials recorded from the round window? *J Acoust Soc Am.* 1998;104(1):370–7. <https://doi.org/10.1121/1.423293>.
- Nakashima T, Pyykko I, Arroll MA, Casselbrant ML, Foster CA, Manzoor NF, et al. Meniere's disease. *Nat Rev Dis Primers.* 2016;2(1):16028. <https://doi.org/10.1038/nrdp.2016.28>.
- Shen D, Wu G, Suk HI. Deep learning in medical image analysis. *Annu Rev Biomed Eng.* 2017;19:221–48. <https://doi.org/10.1146/annurev-bioeng-071516-044442>.
- Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25(1):44–56. <https://doi.org/10.1038/s41591-018-0300-7>.
- Ménière P. Pathologie auriculaire: mémoire sur des lésions de l'oreille interne donnant lieu à des symptômes de congestion cérébrale apoplectiforme. *Gazette médicale de Paris.* 1861;16:597–601.
- Baloh RW. Prosper Meniere and his disease. *Arch Neurol.* 2001;58(7):1151–6. <https://doi.org/10.1001/archneur.58.7.1151>.
- Yamakawa K. Hearing organ of a patient who showed Meniere's symptoms. *J Otolaryngol Soc Jpn.* 1938;44:2310–2.
- Bruderer SG, Bodmer D, Stohler NA, Jick SS, Meier CR. Population-based study on the epidemiology of Meniere's disease. *Audiol Neurotol.* 2017;22(2):74–82. <https://doi.org/10.1159/000475875>.
- Mohamed S, Khan I, Iliodromiti S, Gaggini M, Kontorinis G. Meniere's disease and underlying medical and mental conditions: towards factors contributing to the disease. *ORL J Otorhinolaryngol Relat Spec.* 2016;78(3):144–50. <https://doi.org/10.1159/000444931>.
- Duan M, Bjelke B, Fridberger A, Counter SA, Klason T, Skjonsberg A, et al. Imaging of the Guinea pig cochlea following round window gadolinium application. *Neuroreport.* 2004;15(12):1927–30. <https://doi.org/10.1097/00001756-200408260-00019>.
- Nakashima T, Naganawa S, Sugiura M, Teranishi M, Sone M, Hayashi H, et al. Visualization of endolymphatic hydrops in patients with Meniere's disease. *Laryngoscope.* 2007;117(3):415–20. <https://doi.org/10.1097/MLG.0b013e31802c300c>.
- Naganawa S, Suzuki K, Nakamichi R, Bokura K, Yoshida T, Sone M, et al. Semi-quantification of endolymphatic size on MR imaging after intravenous injection of single-dose gadodiamide: comparison between two types of processing strategies. *Magn Reson Med Sci.* 2013;12(4):261–9. <https://doi.org/10.2463/mrms.2013-0019>.
- Quatre R, Attye A, Karkas A, Job A, Dumas G, Schmerber S. Relationship between audio-vestibular functional tests and inner ear MRI in Meniere's disease. *Ear Hear.* 2019;40(1):168–76. <https://doi.org/10.1097/AUD.0000000000000584>.
- Iida T, Teranishi M, Yoshida T, Otake H, Sone M, Kato M, et al. Magnetic resonance imaging of the inner ear after both intratympanic and intravenous gadolinium injections. *Acta Otolaryngol.* 2013;133(5):434–8. <https://doi.org/10.3109/00016489.2012.753640>.
- Attye A, Eliezer M, Boudiaf N, Tropres I, Chechin D, Schmerber S, et al. MRI of endolymphatic hydrops in patients with Meniere's disease: a case-controlled study with a simplified classification based on saccular morphology. *Eur Radiol.* 2017;27(8):3138–46. <https://doi.org/10.1007/s00330-016-4701-z>.
- Connor SEJ, Pai I. Endolymphatic hydrops magnetic resonance imaging in Meniere's disease. *Clin Radiol.* 2021;76(1):76e1–e19. <https://doi.org/10.1016/j.crad.2020.07.021>.
- Ziyylan F, Smeling DP, Stegeman I, Thomeer HG. Click stimulus electrocochleography versus MRI with intratympanic contrast in Meniere's disease: a systematic review. *Otol Neurotol.* 2016;37(5):421–7. <https://doi.org/10.1097/MAO.0000000000001021>.
- Young YH, Wu CC, Wu CH. Augmentation of vestibular evoked myogenic potentials: an indication for distended saccular hydrops. *Laryngoscope.* 2002;112(3):509–12. <https://doi.org/10.1097/00005537-200203000-00019>.
- Manzari L, Burgess AM, Curthoys IS. Dissociation between cVEMP and oVEMP responses: different vestibular origins of each VEMP? *Eur Arch Otorhinolaryngol.* 2010;267(9):1487–9. <https://doi.org/10.1007/s00405-010-1317-9>.

24. McGarvie LA, Curthoys IS, MacDougall HG, Halmagyi GM. What does the head impulse test versus caloric dissociation reveal about vestibular dysfunction in Meniere's disease? *Ann N Y Acad Sci.* 2015;1343: 58–62. <https://doi.org/10.1111/nyas.12687>.
25. Loureiro RM, Sumi DV, Lemos MD, Tames H, Gomes RLE, Daniel MM, et al. The role of magnetic resonance imaging in Meniere disease: the current state of endolymphatic hydrops evaluation. *Einstein (Sao Paulo).* 2019;17(1):eMD4743. https://doi.org/10.31744/einstein_journal/2019MD4743.
26. Nakashima T, Naganawa S, Pyykko I, Gibson WPR, Sone M, Nakata S, et al. Grading of endolymphatic hydrops using magnetic resonance imaging. *Acta Otolaryngol.* 2009;129(Suppl 560):5–8. <https://doi.org/10.1080/00016480902729827>.
27. Li Z, Wang C, Han M, Xue Y, Wei W, Li L-J et al. Thoracic disease identification and localization with limited supervision. In: 2018 IEEE/CVF conference on computer vision and pattern recognition; 18–23 June 2018. Salt Lake City: IEEE; 2018. p. 8290–9.
28. Fukushima K. Neural network model for a mechanism of pattern recognition unaffected by shift in position-neocognitron. *Trans IECE Jpn A.* 1979;62(10):658–65.
29. Emmert-Streib F, Yang Z, Feng H, Tripathi S, Dehmer M. An introductory review of deep learning for prediction models with big data. *Front Artif Intell.* 2020;3(4) <https://doi.org/10.3389/frai.2020.00004>.
30. Shahamat H, Abadeh MS. Brain MRI analysis using a deep learning based evolutionary approach. *Neural Netw.* 2020;126:218–34. <https://doi.org/10.1016/j.neunet.2020.03.017>.
31. Kong YZ, Gao JL, Xu YP, Pan Y, Wang JX, Liu J. Classification of autism spectrum disorder by combining brain connectivity and deep neural network classifier. *Neurocomputing.* 2019;324:63–8. <https://doi.org/10.1016/j.neucom.2018.04.080>.
32. Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ. Deep learning for brain MRI segmentation: state of the art and future directions. *J Digit Imaging.* 2017;30(4):449–59. <https://doi.org/10.1007/s10278-017-9983-4>.
33. Gürkov R, Jerin C, Flatz W, Maxwell R. Clinical manifestations of hydropic ear disease (Menière's). *Eur Arch Otorhinolaryngol.* 2019;276(1):27–40.
34. Cho YS, Ahn JM, Choi JE, Park HW, Kim YK, Kim HJ, et al. Usefulness of intravenous gadolinium inner ear MR imaging in diagnosis of Meniere's disease. *Sci Rep.* 2018;8(1):17562. <https://doi.org/10.1038/s41598-018-35709-5>.
35. Gurkov R. Meniere and friends: imaging and classification of hydropic ear disease. *Otol Neurotol.* 2017;38(10):e539–e44. <https://doi.org/10.1097/MAO.0000000000001479>.
36. Bouchana A, Kharroubi J, Ridal M. Semi-automatic algorithm for 3D volume reconstruction of inner ear structures based on CT-scan images. In: 2018 4th International conference on advanced technologies for signal and image processing (ATSiP); 21–24 March 2018. Sousse: IEEE; 2018. p. 1–6.
37. Gurkov R, Berman A, Dietrich O, Flatz W, Jerin C, Krause E, et al. MR volumetric assessment of endolymphatic hydrops. *Eur Radiol.* 2015;25(2):585–95. <https://doi.org/10.1007/s00330-014-3414-4>.
38. Wang C, Xi Y. Convolutional neural network for image classification. Baltimore: Johns Hopkins University; 2015. <http://www.cs.jhu.edu/~cwang107/files/cnn.pdf>. Accessed 1 Oct 2020.
39. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal.* 2017;42: 60–88. <https://doi.org/10.1016/j.media.2017.07.005>.
40. Nemzek WR, Brodie HA, Chong BW, Babcock CJ, Hecht ST, Salamat S, et al. Imaging findings of the developing temporal bone in fetal specimens. *AJNR Am J Neuroradiol.* 1996;17(8):1467–77.
41. Cho YS, Cho K, Park CJ, Chung MJ, Kim JH, Kim K, et al. Automated measurement of hydrops ratio from MRI in patients with Meniere's disease using CNN-based segmentation. *Sci Rep.* 2020;10(1):7003. <https://doi.org/10.1038/s41598-020-63887-8>.
42. Fiorino F, Pizzini FB, Barbieri F, Beltramello A. Magnetic resonance imaging fails to show evidence of reduced endolymphatic hydrops in gentamicin treatment of Meniere's disease. *Otol Neurotol.* 2012;33(4): 629–33. <https://doi.org/10.1097/MAO.0b013e318248ee1f>.
43. Wu Q, Dai C, Zhao M, Sha Y. The correlation between symptoms of definite Meniere's disease and endolymphatic hydrops visualized by magnetic resonance imaging. *Laryngoscope.* 2016;126(4):974–9. <https://doi.org/10.1002/lary.25576>.



Jakub Nalepa

Contents

Introduction	1717
Magnetic Resonance Imaging of Brain Tumors	1718
Automated Analysis of MRI in Brain Tumors: Challenges	1718
Structure of the Chapter	1720
Brain Tumor Detection and Segmentation	1720
Automatic Detection and Segmentation of Brain Tumors	1721
Dealing with Limited Ground-Truth Data Sets	1722
Assessing Automatic Detection and Segmentation	1724
Analysis of Segmented Brain Tumors	1725
Quantification of Tumor Characteristics	1725
Classification of Brain Tumors Using AI	1726
Conclusion	1727
References	1728

Abstract

Brain tumors are among the deadliest human cancers, and despite decades of intensive research the survival for many types of malignant primary brain tumors has not improved significantly. Since we continuously generate enormous amounts of clinical data of various modalities that help clinicians not only diagnose brain tumors, but also monitor, quantify, and assess the treatment process, implementing

data-driven approaches to analyze such complex data automatically is becoming extremely important. In this chapter, we review artificial intelligence (AI)-powered approaches for this task, and discuss how AI can bring value into the clinical setting through automating tedious data analysis tasks, and extracting information from medical data that may directly affect the treatment pathway.

Introduction

The amount of medical data generated every second is rapidly growing, and its effective analysis is key for designing an appropriate – and as personalized as possible – patient’s treatment pathway.

J. Nalepa (✉)
Silesian University of Technology, Gliwice, Poland
Future Processing Healthcare, Gliwice, Poland
e-mail: jnalepa@ieee.org; Jakub.Nalepa@polsl.pl

Brain tumors are among the deadliest and feared of all forms of cancer, as more than two-thirds of adults diagnosed with glioblastoma dies within 2 years of diagnosis [3]. Hence, optimizing the clinical workflow and allowing physicians to design the treatment faster and in a reproducible way are critical nowadays – clinical brain tumor imaging plays an integral role in the diagnosis, monitoring, treatment planning, and post-therapy assessment of brain tumors. In this chapter, we discuss how AI helps in the process of tedious analysis of such image data, and ultimately improves the patient care through providing faster, reproducible, and user-independent processing. Additionally, we highlight the most important challenges that have to be faced while designing and implementing the imaging systems benefiting from AI [77].

Magnetic Resonance Imaging of Brain Tumors

Magnetic resonance imaging (MRI) plays a pivotal role in modern brain tumor cancer care because it allows us to non-invasively diagnose a patient, determine the tumor stage, monitor the treatment, assess and quantify its results, and understand its potential side effects. MRI may be exploited to better understand both structural and functional characteristics of the tissue [23]. Such detailed and clinically-relevant analysis of an imaged tumor can ultimately lead to a better patient care. Additionally, MRI does not use the damaging ionizing radiation, and may be utilized to acquire images in different planes and orientations. MRI (with and without contrast) is the investigative tool of choice for neurological cancers – for brain tumors, we acquire multi-modal MRI, including T1-weighted (contrast and non-contrast), T-weighted, Fluid Attenuation Inversion Recovery (FLAIR) sequences, alongside diffusion and perfusion images. For full description of all MRI modalities routinely acquired for brain tumor patients, see an excellent survey by Villanueva-Meyer et al. [89].

An example set of images of various modalities is presented in Fig. 1 (this scan comes from the Multimodal Brain Tumor

Segmentation Challenge [6–9] – we discuss it in detail in section “[Brain Tumor Detection and Segmentation](#)”). We can see that each sequence, rendered in three orientations – axial, coronal, and sagittal – helps evaluate tissue architecture and is captured to manifest its different characteristics:

- **T1** – in pre-contrast sequences, high intensity appears in fat, melanin, blood products, and mineralization, whereas post-contrast T1 is particularly useful in observing the blood-brain barrier and vascular structures. Hence, the post-contrast T1 images may help capture the brain tumor borders, as they become brighter because the contrast agent would accumulate there due to the disruption of the blood-brain barrier in this tumor region [52].
- **T2/FLAIR** – high intensity in T2/FLAIR is seen in peritumoral edema, non-enhancing tumor, as well as gliosis and white matter injury [89].

Automated Analysis of MRI in Brain Tumors: Challenges

Extracting informative features from MRI scans can be represented as a chain of several image analysis and processing tasks (Fig. 2). Importantly, every step directly affects all consecutive blocks in the pipeline. Therefore, any errors or inconsistencies propagate toward the final analysis and interpretation stage. Naturally, each step involves different challenges that need to be faced in order to build a hands-free system that could be deployed in the clinical setting. Here, the segmentation step (i.e., determining different regions of a tumor, e.g., its enhancing part, tumor core, and so forth) is rendered as a dotted block because it may be considered optional – there exist applications in which we are interested in, e.g., quantifying the volume of a whole tumor, not necessarily its specific subparts [65].

The most important challenges concerned with applying AI in brain tumor analysis are summarized in the following bullet points. Note that practically all of them are tightly related to the data availability, quality, and volume:

Fig. 1 Example MRI scan (T2-FLAIR, T2, T1, and post-contrast T1Gd), and the overlaid ground truth (green – peritumoural edema, yellow – enhancing tumor, red – necrotic and non-enhancing tumor core). We can see that different modalities present different tissue characteristics, and the signal intensities captured for the same tumor region substantially vary across them

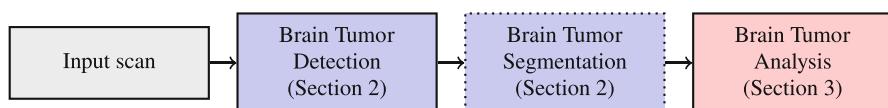
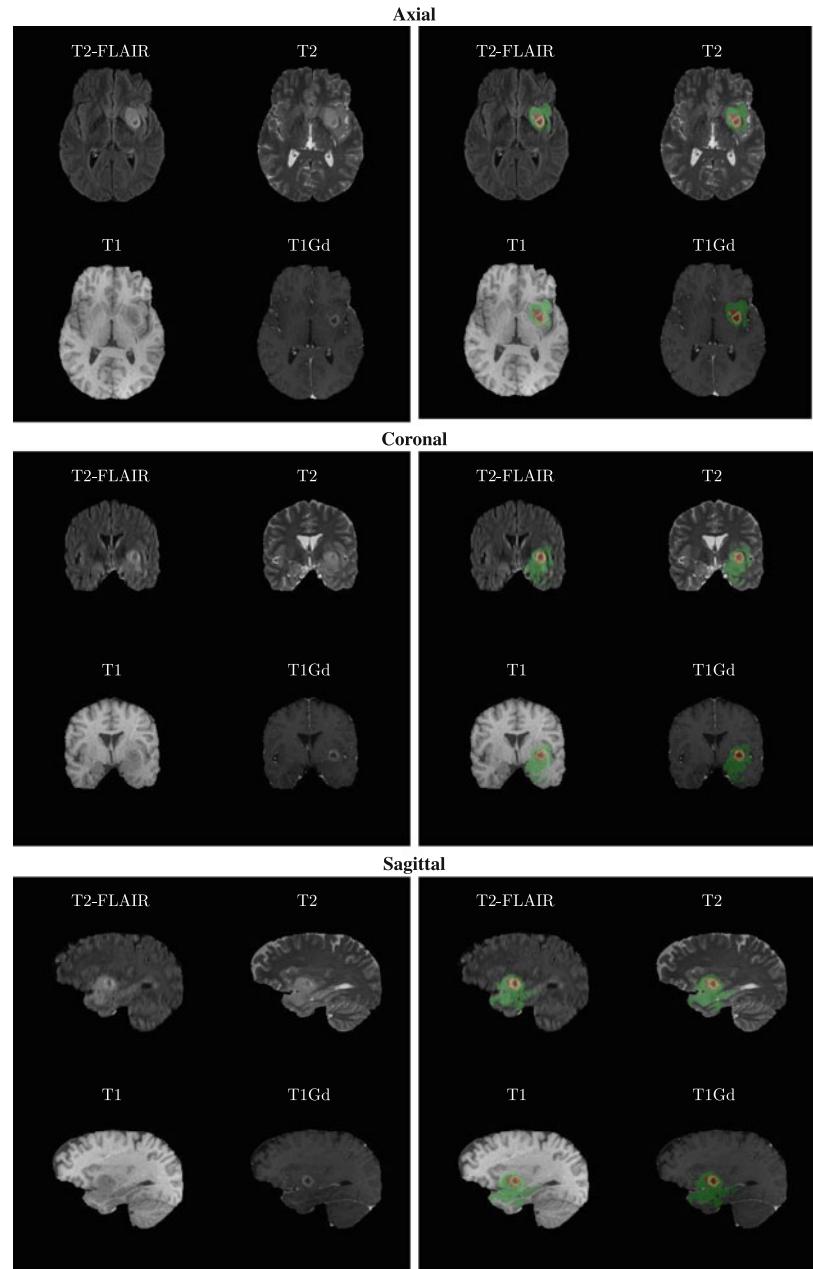


Fig. 2 A high-level flowchart of the automated analysis of brain tumors from MRI. The dotted block presents an optional step of the pipeline that is sometimes omitted in specific use cases

- **The lack of ground-truth data**, being the manually annotated data sets that could be used for training supervised learners. Since

the process of manual contouring of regions of interest in MRI scans is tedious and time-consuming, generating new ground truth is

- often infeasible and cost inefficient. Also, it often involves gathering a group of readers to resolve inter-rater disagreements in the annotation process [65], and to establish “acceptance criteria” for manual delineations.
- **Varying quality of manual delineations in the ground-truth data sets.** Annotating tumor regions is a user-dependent task, and it may be affected not only by a reader’s experience, but also other inter- and intra-operator variability [22]. Therefore, such contouring can easily lead to ground-truth segmentations of questionable quality (hence, it should be manually or semi-automatically improved). Including low-quality image data adversely affects the performance of supervised learners trained over such scans. (However, if the number of such low- or medium-quality examples is small, they could potentially be treated as the data-level regularization [62].)
 - **Very high (often extreme) imbalance of examples** captured in existing ground-truth data sets [46]. It is common that the number of samples (i.e., pixels or voxels) of different classes (e.g., enhancing tumor or necrosis) in a data set is extremely imbalanced, and there exist the majority class(es) dominating the others (called the minority classes). Such data imbalance makes the training process challenging, as it is difficult to capture intrinsic features of the under-represented class(es) from a small number of samples.
 - **Lack of representative ground-truth data sets and the limited representativeness of existing ground truth.** The heterogeneity of healthy brain tissue and brain tumors, alongside the possible locations [72], features, and sizes of tumors make gathering the large and representative training data of brain tumor examples extremely challenging. Also, MRI scans captured using different scanners or acquisition protocols can present significantly different intensity and sequence characteristics – to capture such differences, the ground truth data sets should encompass data acquired using different scanners and protocols [34, 71, 83].
 - **Difficulties in sharing and distributing ground-truth image data across sites (e.g.,**

hospitals). Dealing with medical image data poses new challenges toward its sharing or distributing, especially in the cases where the patient’s information is embedded into the data [91]. Although there exist a plethora of anonymization techniques that help remove “sensitive” information and retain the image data only (without unnecessary metadata), transferring data across sites (and countries) is one of the largest practical challenges nowadays, and there still are situations when it is virtually impossible due to various legal reasons. Therefore, the collaborations between hospitals in the context of building AI-powered systems together may be extremely difficult (or impossible) if they involve data sharing.

Structure of the Chapter

This chapter is structured as follows. In section “[Brain Tumor Detection and Segmentation](#),” we discuss the current advances in the AI-powered brain tumor detection and segmentation, together with the approaches toward dealing with limited ground-truth data, and the ways of assessing the performance of AI models. Section “[Analysis of Segmented Brain Tumors](#)” reviews the most important measures that are typically extracted from brain tumor regions in MRI to quantify the treatment response, perform diagnosis (or prognosis), and better understand the effectiveness of the undertaken treatment pathway. Finally, section “[Conclusion](#)” concludes the chapter.

Brain Tumor Detection and Segmentation

Detection and segmentation of brain tumors from MRI are the critical steps in a process of analyzing an MRI study, as they significantly influence, e.g., extracting quantifiable tumor characteristics. Here, by *detection* we mean the process of determining tumorous pixels (or voxels in 3D) in the input scan, and by *segmentation* – classifying such pixels/voxels into specific tumor subparts (e.g., edema, enhancing part of a tumor, tumor core, and so on).

Incorrect delineation may easily lead to improper interpretation of the captured scan and thus can adversely affect the treatment pathway [59].

A tremendous amount of MRI data generated every day drives the development of machine learning-powered brain-lesion segmentation systems. However, such data is extremely imbalanced (only the minority of all pixels or voxels are tumorous), it is very large (in terms of volume) and heterogeneous. This heterogeneity may be not only due to different scanners and/or protocols utilized to capture MRI data, but may also be a result of various tumor characteristics and intrinsic features captured in the image. Therefore, manual delineation of MRI scans is time-consuming and tedious, and generating high-quality manually-segmented brain lesions (that could be exploited for training) is challenging in practice. It is worth mentioning that there might be discrepancies and disagreements between readers delineating the very same scan (or even between two segmentations provided by the same reader at different time points), and – in general – the quality of manual segmentations vary [90].

Automatic Detection and Segmentation of Brain Tumors

Fully-automated medical-image segmentation pipelines, e.g., exploiting image analysis and machine learning, are of great interest, as they can accelerate diagnosis, ensure reproducibility, and make comparisons much easier (e.g., comparing the manual delineations performed by two readers at two different oncology centers alongside the extracted brain-tumor numerical features is extremely challenging if they did not follow the same acquisition and segmentation protocol). Also, we are aimed at decreasing the overall analysis

time. Finally, we want to reduce false negatives, being the tumorous voxels that were incorrectly classified as healthy by an algorithm or a human.

In Fig. 3, we present a high-level taxonomy of the algorithms for automated detection and segmentation of brain tumors from MRI. In the *atlas-based* techniques, manually segmented images (*atlases*) that model anatomical variability of the brain are exploited to segment unseen scans [66, 70]. Thus, to ensure that such algorithms can be successfully applied to new MRI scans, creating large and representative annotated sets is pivotal yet very costly [4]. Additionally, atlas approaches are dependent on the quality of the underlying image registration process [11]. The *image analysis*-powered algorithms are commonly split into *thresholding-* and *region-based* techniques [40]. The former algorithms extract the threshold values (either single or multiple of them, potentially in an adaptive way) to classify voxels into different classes, whereas the latter utilize the neighborhood pixel information (quantified using a selected criteria) during the segmentation process [21]. Other techniques that benefit from image analysis include active contours (deformable models) [12], region-based active contours [92], and elastic image deformations [63].

The *classical machine learning* brain tumor detection and segmentation algorithms involve feature engineering (i.e., feature extraction, commonly followed by feature selection), whereas *deep learning* techniques allow us to perform automatic representation learning. Both approaches can be further divided into *unsupervised* and *supervised* algorithms. The former operate on the unlabeled data and reveal its hidden structures and characteristics through, e.g., clustering [15, 25, 58, 86]. *Supervised* techniques use manually delineated scans as examples to train a model, and include a plethora of techniques, such as

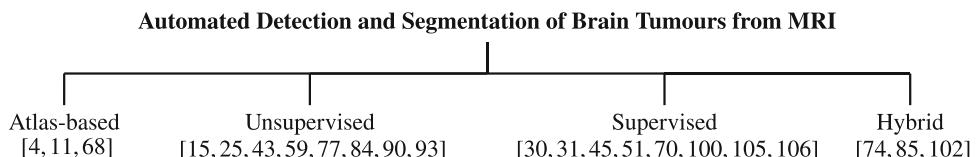


Fig. 3 Automated delineation of brain tumors from MRI – a taxonomy [56]

decision forests [30, 101], conditional random fields [95], extremely randomized forests [69], support vector machines [49], k -nearest neighbors [44, 82], various architectures of artificial neural networks [1], and many more. We have been witnessing the unprecedented success of deep learning in practically all areas of science and industry, and medical image analysis is not an exception here. Deep networks have been applied to process brain scans [31, 37, 61, 100] – they consistently outperform other approaches in the Multimodal Brain Tumor Segmentation Challenge (BraTS) [43]. We can observe that the U-Net-shaped architectures (in both 2D and 3D) are currently the “architecture of choice” in automated brain tumor delineation [41]. Although there exist a multitude of variants of the vanilla U-Net [19, 26, 56, 87], the appropriate pre- and post-processing, alongside the configuration of the basic architecture are key in obtaining very accurate detection and segmentation [42]. Finally, *hybrid* approaches combine methods from other categories [81, 97].

Although multi-modal processing is pivotal for segmenting brain tumors into sub-regions, they can be effectively detected (without dividing them into sub-parts) from a single MRI modality, e.g., T2-FLAIR [54, 65, 89, 96]. This approach is often much easier to implement, train, and deploy, and infers in shorter time, especially due to the lack of the sequence co-registration step required in multi-modal techniques. Additionally, it is common that some modalities are missing in clinical scenarios, hence designing and implementing brain tumor analysis algorithms robust against the absence of a modality (or a set of modalities), and quantifying their influence on the overall segmentation quality are of research interest [80, 98].

Automatic brain tumor detection and segmentation has become a mature and very active research topic – a battery of available approaches allow the practitioners select and implement such methods in clinical settings. Although the progress in the field has been driven by the revolutionizing impact of deep learning, all of the aforementioned techniques have their own advantages and shortcomings, and are often applied in different use cases. For a comprehensive review of the state of the art,

coupled with a detailed analysis of their pros and cons, see [14, 65, 76, 92].

Dealing with Limited Ground-Truth Data Sets

Questionable quality of manually-generated ground-truth information is an important problem in the case of supervised methods, because it directly affects the abilities of a trained model. Additionally, gathering well-represented heterogeneous MRI examples, i.e., scans acquired using a variety of MR scanners and imaging protocols, is extremely costly and time-consuming. In BraTS, the authors publish a set of MRI scans (obtained at 19 institutions) – the T2-FLAIR, T2, T1, and T1c (T1Gd) sequences are co-registered, skull-stripped, and interpolated to the resolution of 1 mm^3 (155 frames of 240×240 size) [6–9]. The scans are coupled with manual annotations generated by one to four experienced readers who followed the same delineation protocol. Afterwards, their annotations were double-checked by neuro-radiologists, and the following classes of voxels were included in the data set: GD-enhancing tumor, the peritumoral edema, and the necrotic and non-enhancing tumor core. In the newest (2020) release of the challenge, the training set contained 293 high-grade glioblastomas and 76 low-grade gliomas with manual annotations, and – to the best of our knowledge – is the largest and the most comprehensive data set that can be used for training supervised learners, and for quantifying the performance of existing and emerging algorithms.

Although the BraTS data set is widely used in the field, the number of patients included in this set is still relatively small and dealing with limited (in terms of the size, heterogeneity or specific-class examples) ground-truth data sets remains a valid real-life challenge. The approaches toward tackling this issue may be divided into:

- **Unsupervised and semi-supervised learning.**
As discussed earlier in this section, unsupervised methods do not require any ground-truth data to uncover “hidden” features of the raw

data. Although such techniques allow us to determine consistent 2D/3D regions that manifest similar characteristics, we miss the class label once the pixels/voxels are grouped. Thus, such pre-processed scans undergo additional analysis to, e.g., manually classify the segmented regions. Importantly, unsupervised learning may accelerate the process of generating ground truth through elaborating pre-segmented areas that are later classified into the classes by a human. On the other hand, semi-supervised algorithms make full use of the unlabeled data (note that a multitude of unlabeled MRI scans are generated daily worldwide), and they estimate the missing labels. Then, these samples can be used together with the labelled data for training a supervised learner [29].

- **Transfer learning and domain adaptation.** In transfer learning (Fig. 4), we exploit existing ground-truth data sets (referred to as the *source data sets* in this context), very often captured in different domains, to train a deep learning algorithm in a supervised way. Then, the feature extractor is transferred to the target domain, and the classification part of the model is further trained over the *target data set* (of a much smaller size). In this scenario,

we still need the target (labelled) data but it may be of a significantly smaller size.

- **Synthesizing artificial training examples through data augmentation.** Data augmentation is a process of generating synthetic samples [62]. Such artificial images can be either included in the training set (in order to increase its size and representativeness) in the *training-time* augmentation, or utilized during the inference to benefit from the ensemble-like approach in the *test-time* augmentation approach. In the latter case, we generate n synthetic samples for each original input sample, classify ($n-1$) samples, and aggregate the classification results (e.g., via the majority voting) into the final label assigned to the original sample. Data augmentation approaches are divided into the techniques that operate on the existing data via applying simple image transforms such as rotation, flipping, shearing, translation and zooming [28, 53], alongside elastic [13, 33, 38, 63] or pixel-level transformations [2, 75], and those that synthesize new examples, using, e.g., generative adversarial networks [99] or various tumor growth models [32] (Fig. 5). The former techniques often generate correlated image samples but can be employed at test time, as they can synthesize

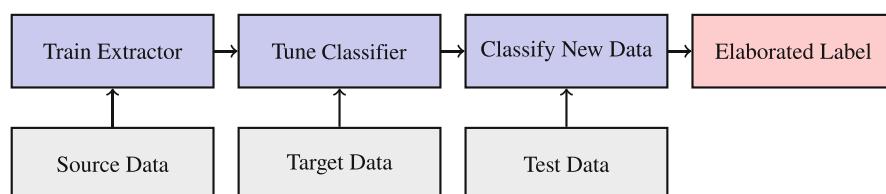


Fig. 4 In transfer learning, we train feature extractors over the source data, and fine-tune the classifier over the target data. This figure is inspired by [64]

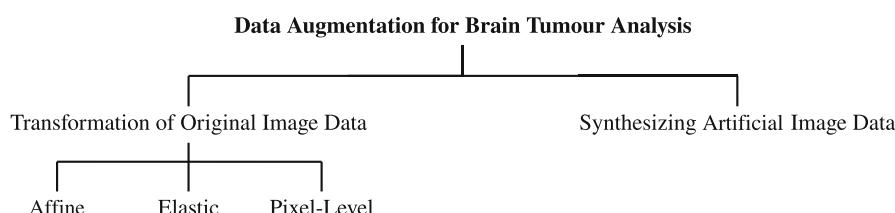


Fig. 5 Data augmentation for brain-tumor detection and segmentation – a taxonomy [62]

artificial images using an incoming test image. On the other hand, the methods that generate image data with the use of generative models can only be deployed before training the models to enhance the training set. Data augmentation is a standard procedure that proved to be pivotal in obtaining well-generalizing models, and it is consistently exploited in the best-performing tumor detection and segmentation systems [62].

- **Federated learning.** Multi-institutional collaborations in the context of medical image analysis are extremely difficult in the scenarios involving centralized data centers, in which data would be shared across sites, due to the issues that are concerned with data anonymization, privacy, and ownership. To leverage the unprecedented amount of data that can be captured at numerous sites (clinics, universities, research centers, and so forth), federated learning – as a new distributed training paradigm – can be utilized [78]. Here, the process of training the models is distributed across the collaborators, and each collaborator trains the model using their data, hence the data privacy is fully maintained. Afterwards, the models are “aggregated” and assessed before the final deployment [79].

Assessing Automatic Detection and Segmentation

Quantifying the performance of automatic brain tumor detection and segmentation methods is critical to fully understand their performance abilities.

In general, we distinguish three ways of assessing the quality of such algorithms that are commonly used to thoroughly analyze emerging techniques:

- **Quantitative analysis.** To quantitatively measure the quality of automatic segmentations, we utilize various metrics extracted from the confusion matrix containing true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). They commonly include the spatial overlap metrics: DICE coefficient (sometimes coupled with the Jaccard’s index; both metrics range from zero to one, and both should be maximized: ↑ – see an example rendered in Fig. 6), true positive rate (sensitivity, recall; ranging from zero to one, ↑), true negative rate (specificity; ranging from zero to one, ↑), and other related metrics [85]. On the other hand, the spatial distance metrics quantify the quality of the boundary (contour) [27], and encompass the Hausdorff distance (↓), Mahalanobis distance (↓), together with their variants, usually more robust to outlying (low-quality) delineations [39]. It is worth noting that the aforementioned metrics are calculated with respect to the ground-truth segmentations. Therefore, the quality of ground truth directly influences the extracted metrics, and – in extreme scenarios – low values of, e.g., the overlap metrics do not necessarily mean that the automatically contoured regions are of poor quality, but they may indicate the poor quality of reference annotations delivered, e.g., by inexperienced readers. Hence, the quantitative analysis should be coupled with the qualitative analysis that could reduce

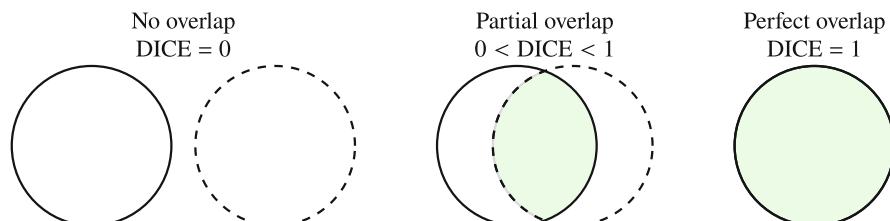


Fig. 6 Example values of the DICE scores obtained for two contours annotated using solid and dashed lines. The larger overlapping areas are reflected in greater DICE values

the impact of “misleading” quantitative metrics obtained for ground truth of questionable or varying quality.

- **Qualitative analysis.** Exploring and investigating the automatic delineations is pivotal not only while developing the algorithms, but also during their verification and validation. In [65], the authors exploited the mean opinion score experiment as a framework for assessing the quality of automatically determined brain tumor contours – a group of readers were asked to investigate the annotations elaborated using a deep model, and score their clinical utility using a simple scale (“very low-quality” and “low-quality” segmentation that map to “I would not use this segmentation to support diagnosis,” and “acceptable” and “high-quality” segmentation that map to “I would use this segmentation to support diagnosis”). Since the readers may easily be biased, the authors analyzed inter-rater agreements and delivered secondary analysis over the results of this experiment. Such qualitative investigation and assessment of image data performed by human practitioners is extremely important, as it can prove the clinical utility of a developed algorithm (or show that the results are not mature enough to be deployed in a clinical setting).
- **Statistical analysis.** To fully understand the statistical relevance of the obtained results (together with the differences between the investigated variants of the segmentation techniques), we commonly execute statistical testing [102]. It may be exploited to, e.g., verify the null hypothesis saying that “applying Algorithm A and Algorithm B leads to segmentations of the same quality, quantified, e.g., using the DICE score.” Also, as already mentioned, investigating the inter-rater agreement is critical to demonstrate consistency among observational ratings provided by multiple readers [48, 90], but – at the same time – incorrect statistical procedures are commonly utilized in this context [35]. Statistical tools are also used to prove the reproducibility of the results delivered by AI algorithms [67, 74].

Analysis of Segmented Brain Tumors

AI algorithms are often utilized to automate brain tumor detection and segmentation, but these are commonly the intermediate steps in a processing chain. In this section, we discuss two important analysis steps that may follow the process of locating the tumors: quantifying tumor characteristics and features (section “[Quantification of Tumor Characteristics](#)”), and classifying delineated tumors into specific classes (section “[Classification of Brain Tumors Using AI](#)”).

Quantification of Tumor Characteristics

Currently, the most widely-used criteria of assessing response to therapy in high-grade brain tumors are based on single- or two-dimensional measurements of an enhancing part of a tumor in computed tomography or MRI [94]. Although they have limitations, as enhancement is non-specific, hence may not always be an appropriate surrogate of tumor response, and some therapies result in transient increase of tumor enhancement (referred to as the pseudo-progression), such criteria are consistently utilized in clinical practice to analyze longitudinal scans. In Table 1, we gather the popular criteria, alongside their basic features, including the image modality, the way of determining target lesions, and the shrinkage required for partial progression [51]. Importantly, these response criteria are the result of the international effort of building a standardized set of measures that allow us for designing the objective assessment of treatment response. Also, they are commonly reported together with other volumetric and bidimensional measurements extracted for the analyzed tumors. Establishing such criteria is a dynamic process, as we gather new underpinnings of primary and metastatic brain tumors, therefore new or updated measures are likely to emerge in the nearest future [18].

Because the response criteria are based on one- or two-dimensional measurements of a tumor, determining such measures involves detecting and segmenting of a lesion in an input scan – see an example of visualized diameters that are used to calculate the Response Assessment in Neuro-

Table 1 The criteria commonly used to assess response to therapy in brain tumor patients, alongside the shrinkage required for partial response. If a patient has multiple

lesions, the enlarging lesions are considered as target lesions. For more details, see [51]

Criterion	Modality	Target lesion	Measurement type	Shrinkage (%)
RECIST 1.0 [88]	MRI or CT	Longest diameter >10 mm	Unidimensional	≥ 30
RECIST 1.1 [24]	MRI or CT	Longest diameter >10 mm	Unidimensional	≥ 30
Macdonald [55]	MRI or CT	Minimal size is not specified	Bidimensional	≥ 50
WHO [60]	–	Minimal size is not specified	Bidimensional	≥ 50
RANO [94]	MRI or CT	Contrast enhancing lesions with two perpendicular diameters (> 10 mm)	Bidimensional	≥ 50

Oncology (RANO) criterion presented in Fig. 7. Thus, AI techniques for these tasks are naturally embedded into the systems that are designed to extract quantifiable tumor measures utilized to track treatment response [16] (note that elaborating, e.g., the perpendicular diameters in RANO does not involve any AI and is a pure image analysis and optimization task). Finally, quantifying the impact of automatic segmentation quality on the biomarkers extracted from MR brain imaging is a vital research topic. In [65], Nalepa et al. implemented a fully-automated deep learning-powered technique for extracting biomarkers from dynamic contrast-enhanced MRI (DCE-MRI) of brain tumors. Once the contrast bolus is injected, DCE-MRI captures the voxel intensity changes within a volume of interest, and it allows for quantifying the dynamic processes within a tissue based on such temporal changes of voxel intensities. Biomarkers extracted from DCE-MRI can be used in patient prognosis, risk assessment and quantification of tumor characteristics and stage – such analysis of the temporal enhancement pattern can benefit from AI for segmenting the tumorous regions as well.

Classification of Brain Tumors Using AI

Many computer-aided diagnosis systems are aimed at not only detecting and segmenting brain tumors from MRI, but also at classifying them into specific classes. Such classes may reflect their

aggressiveness, stage, or type, e.g., benign and malignant; primary and secondary; grade II, III, and IV tumors; glioma, meningioma, and pituitary tumor [45]. Machine learning algorithms for brain tumor classification (both two- and multi-class) include classical techniques that involve feature extraction, selection, and training, alongside deep learning methods that automate feature extraction. Handcrafted features commonly include texture, statistical, shape, and grey-level characteristics of the tumorous part of the brain [36, 73] – such extracted features form feature vectors that are fed into the training process of a supervised learner, e.g., a support vector machine [5] (Fig. 8). On the other hand, deep learning models embed the feature extractor into the network architecture and make it possible to train [84, 93]. As an example, trainable kernels in convolutional neural networks act as such feature extractors. There exist hybrid approaches [47], in which hand-crafted tumor features are coupled with the deep ones to fully exploit both the “expert knowledge” that may be manifested in manually designed extractors, and data-driven feature extractors (e.g., autoencoders or convolutional feature extractors) that may potentially capture features that can be “unknown” for humans [10]. Since the issues concerned with the limited number of training examples are valid in the context of supervised tumor classification, the same approaches are utilized to tackle them [17, 20, 57], as in the segmentation task (see section “Dealing with Limited Ground-Truth Data Sets” for details).

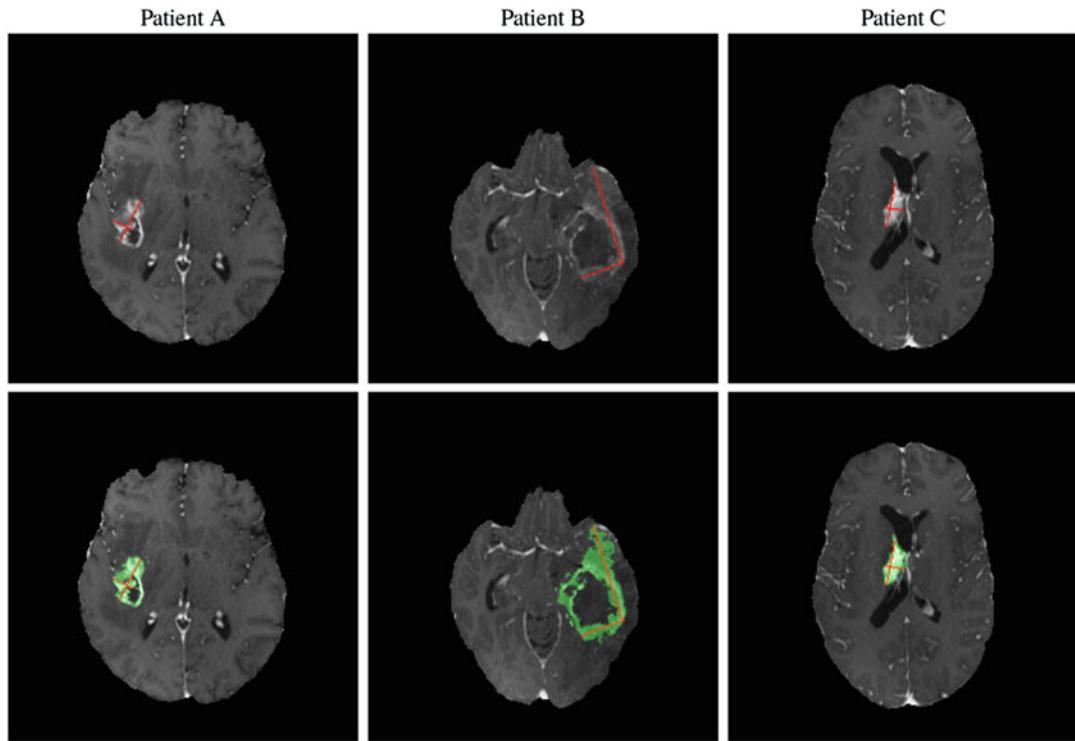


Fig. 7 Example BraTS brain tumor scans with annotated enhancing tumor (in green), alongside the perpendicular diameters (in red) used for calculating RANO, being their product

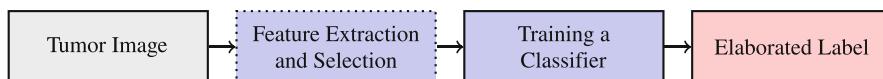


Fig. 8 Classification of brain tumors using classical machine learning (with manually designed feature extractors and selectors, annotated as a dotted block) and deep learning

Conclusion

AI algorithms are being continuously deployed in medical data analysis systems, and can fundamentally affect different levels of the analysis workflow (e.g., enhancing the quality of medical images, improving their interpretation, automating tedious analysis tasks, and extracting mineable high-dimensional data to enhance clinical decision making, just to mention a few workflow steps that could easily benefit from AI). Although it is tempting to think that AI can effectively solve all emerging problems in automated analysis of high-dimensional medical data, there are a

multitude of challenges that have to be tackled before deploying such data-driven techniques in clinical practice [68]. In this chapter, we discussed how AI can improve the brain tumor patients' care through automating pivotal steps of analyzing image data. We identified the most critical obstacles that are faced while designing, verifying, and validating new AI algorithms, and presented the ways of dealing with them.

AI will transform (and it actually *is already transforming*) clinical imaging practice [50]. It is, however, important to realize that AI *is not* to replace clinicians, but to deliver tools that can automate their most tedious, time-consuming,

and user-dependent tasks, and to help them move the focus to the actual interpretation of the results and designing more personalized and even better treatment pathways.

Acknowledgments This chapter is in memory of Dr. Grzegorz Nalepa, an extraordinary scientist and pediatric hematologist/oncologist at Riley Hospital for Children, Indianapolis, USA, who helped countless patients and their families through the most challenging moments of their lives.

The work was supported by the Silesian University of Technology grant for maintaining and developing research potential, and by Rectors Research and Development Grant 02/080/RGJ20/0003. The author thanks Krzysztof Kotowski for implementing the RANO visualization.

References

1. Abdalla HEM, Esmail MY. Brain tumor detection by using artificial neural network. In: Proceedings of ICCCEEE; 2018. p. 1–6.
2. Agarwal M, Mahajan R. Medical images contrast enhancement using quad weighted histogram equalization with adaptive GAMA correction and homomorphic filtering. Procedia Comput Sci. 2017;115: 509–17.
3. Aldape K, Brindle KM, Chesler L, Chopra R, Gajjar A, Gilbert MR, Gottardo N, Gutmann DH, Hargrave D, Holland EC, Jones DTW, Joyce JA, Kearns P, Kieran MW, Mellinghoff IK, Merchant M, Pfister SM, Pollard SM, Ramaswamy V, Rich JN, Robinson GW, Rowitch DH, Sampson JH, Taylor MD, Workman P, Gilbertson RJ. Challenges to curing primary brain tumours. Nat Rev Clin Oncol. 2019;16(8):509–20.
4. Aljabar P, Heckemann R, Hammers A, Hajnal J, Rueckert D. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. NeuroImage. 2009;46(3):726–38.
5. Ansari MA, Mehrotra R, Agrawal R. Detection and classification of brain tumor in MRI images using wavelet transform and support vector machine. J Interdiscip Math. 2020;23(5):955–66.
6. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby J, Freymann J, Farahani K, Davatzikos C. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Nat Sci Data. 2017a;4:1–13. <https://doi.org/10.1038/sdata.2017.117>.
7. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby JS, Freymann JB, Farahani K, Davatzikos C. Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. Cancer Imaging Arch. 2017b. <https://doi.org/10.7937/K9/TCIA.2017.KLXWJJ1Q>.
8. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby JS, Freymann JB, Farahani K, Davatzikos C. Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. Cancer Imaging Arch. 2017c. <https://doi.org/10.7937/K9/TCIA.2017.GJQ7R0EF>.
9. Bakas S, Reyes M, Jakab A, Bauer S, Remper M, Crimi A, Shinohara RT, Berger C, Ha SM, Rozycki M, Prastawa M, Alberts E, Lipková J, Freymann JB, Kirby JS, Bilello M, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. CoRR abs/1811.02629. 2018. <http://arxiv.org/abs/1811.02629>
10. Basheera S, Ram MSS. Classification of brain tumors using deep features extracted using CNN. J Phys Conf Ser. 2019;1172:012016.
11. Bauer S, Seiler C, Bardyn T, Buechler P, Reyes M. Atlas-based segmentation of brain tumor images using a Markov Random Field-based tumor growth model and non-rigid registration. In: Proceedings of IEEE EMBC; 2010. p. 4080–3.
12. Ben Rabeh A, Benzarti F, Amiri H. Segmentation of brain MRI using active contour model. Int J Imaging Syst Technol. 2017;27(1):3–11.
13. Castro E, Cardoso JS, Pereira JC. Elastic deformations for data augmentation in breast cancer mass detection. In: Proceedings of IEEE BHI; 2018. p. 230–4.
14. Chahal PK, Pandey S, Goel S. A survey on brain tumor detection techniques for MR images. Multimed Tools Appl. 2020;79(29):21771–814.
15. Chander A, Chatterjee A, Siarry P. A new social and momentum component adaptive PSO algorithm for image segmentation. Expert Syst Appl. 2011;38(5): 4998–5004.
16. Chang K, Beers AL, Bai HX, Brown JM, Ly KI, Li X, Senders JT, Kavouridis VK, Boaro A, Su C, Bi WL, Rapalino O, Liao W, Shen Q, Zhou H, Xiao B, Wang Y, Zhang PJ, Pinho MC, Wen PY, Batchelor TT, Boxerman JL, Arnaout O, Rosen BR, Gerstner ER, Yang L, Huang RY, Kalpathy-Cramer J. Automatic assessment of glioma burden: a deep learning algorithm for fully automated volumetric and bidimensional measurement. Neuro-Oncology. 2019;21(11):1412–22.
17. Cheng J, Huang W, Cao S, Yang R, Yang W, Yun Z, Wang Z, Feng Q. Enhanced performance of brain tumor classification via tumor region augmentation and partition. PLoS One. 2015;10(10):1–13.
18. Chukwueke UN, Wen PY. Use of the response assessment in neuro-oncology (RANO) criteria in clinical trials and clinical practice. CNS Oncol. 2019;8(1): CNS28.
19. Dai L, Li T, Shu H, Zhong L, Shen H, Zhu H. Automatic brain tumor segmentation with domain adaptation. In: Crimi A, Bakas S, Kuijf H, Keyvan F, Reyes M, van Walsum T, editors. Brainlesion: glioma,

- multiple sclerosis, stroke and traumatic brain injuries. Cham: Springer International Publishing; 2019. p. 380–92.
- 20. Deepak S, Ameer P. Brain tumor classification using deep CNN features via transfer learning. *Comput Biol Med*. 2019;111:103345.
 - 21. Deng W, Xiao W, Deng H, Liu J. MRI brain tumor segmentation with region growing method based on the gradients and variances along and inside of the boundary curve. In: Proceedings of ICBEI, vol. 1; 2010. p. 393–6.
 - 22. Despotović I, Goossens B, Philips W. MRI segmentation of the human brain: challenges, methods, and applications. *Computat Math Methods Med*. 2015;2015:450341.
 - 23. Dickie DA, Shenkin SD, Anblagan D, Lee J, Blesa Cabez M, Rodriguez D, Boardman JP, Waldman A, Job DE, Wardlaw JM. Whole brain magnetic resonance image atlases: a systematic review of existing atlases and caveats for use in population imaging. *Front Neuroinform*. 2017;11:1.
 - 24. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, Dancey J, Arbuck S, Gwyther S, Mooney M, Rubinstein L, Shankar L, Dodd L, Kaplan R, Lacombe D, Verweij J. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer*. 2009;45(2):228–47.
 - 25. Fan X, Yang J, Zheng Y, Cheng L, Zhu Y. A novel unsupervised segmentation method for MR brain images based on fuzzy methods. In: Liu Y, Jiang T, Zhang C, editors. *Proceedings of CVBIA*. Berlin: Springer; 2005. p. 160–9.
 - 26. Fang L, He H. Three pathways U-Net for brain tumor segmentation. In: *Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries – 4th international workshop, BrainLes 2018, held in conjunction with MICCAI 2018, Granada, Spain, pre-conference proceedings*; 2018. p. 119–26.
 - 27. Fenster A, Chiu B. Evaluation of segmentation algorithms for medical imaging. In: *Proceedings of IEEE EMB*; 2005. p. 7186–9.
 - 28. Frid-Adar M, Diamant I, Klang E, Amitai M, Goldberger J, Greenspan H. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*. 2018;321:321–31.
 - 29. Ge C, Gu IYH, Jakola AS, Yang J. Deep semi-supervised learning for brain tumor classification. *BMC Med Imaging*. 2020;20(1):87.
 - 30. Geremia E, Clatz O, Menze BH, Konukoglu E, Criminisi A, Ayache N. Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. *NeuroImage*. 2011;57(2):378–90.
 - 31. Ghafoorian M, Mehrtash A, Kapur T, Karssemeijer N, Marchiori E, Pesteie M, Guttmann CRG, de Leeuw FE, Tempany CM, van Ginneken B, Fedorov A, Abolmaesumi P, Platel B, Wells W. Transfer learning for domain adaptation in MRI: application in brain lesion segmentation. In: *Proceedings of MICCAI*; 2017. p. 516–24.
 - 32. Gholami A, Subramanian S, Shenoy V, Himthani N, Yue X, Zhao S, Jin PH, Biros G, Keutzer K. A novel domain adaptation framework for medical image segmentation. In: *Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries – 4th international workshop, BrainLes 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, revised selected papers, part II*; 2018. p. 289–98.
 - 33. Gu S, Meng X, Sciurba FC, Ma H, Leader JK, Kaminski N, Gur D, Pu J. Bidirectional elastic image registration using B-spline affine transformation. *Comput Med Imaging Graph*. 2014;38(4):306–14.
 - 34. Guo C, Niu K, Luo Y, Shi L, Wang Z, Zhao M, Wang D, Zhu W, Zhang H, Sun L. Intra-scanner and inter-scanner reproducibility of automatic white matter hyperintensities quantification. *Front Neurosci*. 2019;13:679.
 - 35. Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutor Quant Methods Psychol*. 2012;8(1):23–34.
 - 36. Hamid MAA, Khan NA. Investigation and classification of MRI brain tumors using feature extraction technique. *J Med Biol Eng*. 2020;40(2):307–17.
 - 37. Havaei M, Davy A, Warde-Farley D, Biard A, Courville A, Bengio Y, Pal C, Jodoin PM, Larochelle H. Brain tumor segmentation with deep neural networks. *Med Image Anal*. 2017;35:18–31.
 - 38. Huang Z, Cohen FS. Affine-invariant B-spline moments for curve matching. *IEEE Trans Image Process*. 1996;5(10):1473–80.
 - 39. Huttenlocher DP, Klanderman GA, Rucklidge WJ. Comparing images using the Hausdorff distance. *IEEE Trans Pattern Anal Mach Intell*. 1993;15(9):850–63.
 - 40. Ilhan U, Ilhan A. Brain tumor segmentation based on a new threshold approach. *Procedia Comput Sci*. 2017;120:580–7.
 - 41. Isensee F, Kickingereder P, Wick W, Bendszus M, Maier-Hein KH. No newnet. In: Crimi A, Bakas S, Kuijf H, Keyvan F, Reyes M, van Walsum T, editors. *Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries*. Cham: Springer International Publishing; 2019. p. 234–44.
 - 42. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2020;18:203.
 - 43. Kamnitsas K, Bai W, Ferrante E, McDonagh SG, Sinclair M, Pawlowski N, Rajchl M, Lee M, Kainz B, Rueckert D, Glocker B. Ensembles of multiple models and architectures for robust brain tumour segmentation. In: *Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injury*. Springer; 2018. p. 450–62.
 - 44. Khalid N, Ibrahim S, Haniff P. MRI brain abnormalities segmentation using k-NN. *Int J Comput Sci Eng*. 2011;3(2):980–90.

45. Khan MA, Ashraf I, Alhaisoni M, Damaevius R, Scherer R, Rehman A, Bukhari SAC. Multimodal brain tumor classification using deep learning and robust feature selection: a machine learning application for radiologists. *Diagnostics*. 2020;10(8):565.
46. Krawczyk B. Learning from imbalanced data: open challenges and future directions. *Prog Artif Intell*. 2016;5(4):221–32.
47. Kumar S, Dabas C, Godara S. Classification of brain MRI tumor images: a hybrid approach. *Procedia Comput Sci*. 2017;122:510–7.
48. Kvåleseth TO. Measurement of interobserver disagreement: correction of Cohen's kappa for negative values. *J Probab Stat*. 2015;2015:751803.
49. Ladgham A, Torkhani G, Sakly A, Mtibaa A. Modified support vector machines for MR brain images recognition. In: Proceedings of CoDIT; 2013. p. 032–5.
50. Langlotz CP, Allen B, Erickson BJ, Kalpathy-Cramer J, Bigelow K, Cook TS, Flanders AE, Lungren MP, Mendelson DS, Rudie JD, Wang G, Kandarpa K. A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 NIH/RSSNA/ACR/The Academy Workshop. *Radiology*. 2019;291(3):781–91.
51. Lin NU, Lee EQ, Aoyama H, Barani IJ, Barboriak DP, Baumert BG, Bendszus M, Brown PD, Camidge DR, Chang SM, Dancey J, de Vries EGE, Gaspar LE, Harris GJ, Hodis FS, Kalkanis SN, Linskey ME, Macdonald DR, Margolin K, Mehta MP, Schiff D, Soffietti R, Suh JH, van den Bent MJ, Vogelbaum MA, Wen PY. Response assessment criteria for brain metastases: proposal from the RANO group. *Lancet Oncol*. 2015;16(6):e270–8.
52. Liu J, Li M, Wang J, Wu F, Liu T, Pan Y. A survey of MRI-based brain tumor segmentation methods. *Tsinghua Sci Technol*. 2014;19(6):578–95.
53. Liu Y, Stojadinovic S, Hrycushko B, Wardak Z, Lau S, Lu W, Yan Y, Jiang SB, Zhen X, Timmerman R, Nedzi L, Gu X. A deep convolutional neural network-based automatic delineation strategy for multiple brain metastases stereotactic radiosurgery. *PLoS One*. 2017;12(10):1–17.
54. Lorenzo PR, Nalepa J, Bobek-Billewicz B, Wawrzyniak P, Mrukwa G, Kawulok M, Ulrych P, Hayball MP. Segmenting brain tumors from FLAIR MRI using fully convolutional neural networks. *Comput Methods Prog Biomed*. 2019;176:135–48.
55. Macdonald DR, Cascino TL, Schold SC, Cairncross JG. Response criteria for phase II studies of supratentorial malignant glioma. *J Clin Oncol*. 1990;8(7):1277–80.
56. Marcinkiewicz M, Nalepa J, Lorenzo PR, Dudzik W, Mrukwa G. Segmenting brain tumors from MRI using cascaded multi-modal U-Nets. In: Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries – 4th international workshop, BrainLes 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, revised selected papers, part II; 2018. p. 13–24.
57. Mehrotra R, Ansari M, Agrawal R, Anand R. A transfer learning approach for AI-based classification of brain tumors. *Mach Learn Appl*. 2020;2:100003.
58. Mei PA, de Carvalho Carneiro C, Fraser SJ, Min LL, Reis F. Analysis of neoplastic lesions in magnetic resonance imaging using self-organizing maps. *J Neurol Sci*. 2015;359(1–2):78–83.
59. Meier R, Knecht U, Loosli T, Bauer S, Slotboom J, Wiest R, Reyes M. Clinical evaluation of a fully-automatic segmentation method for longitudinal brain tumor volumetry. *Sci Rep*. 2016;6(1):23376.
60. Miller AB, Hoogstraten B, Staquet M, Winkler A. Reporting results of cancer treatment. *Cancer*. 1981;47(1):207–14.
61. Moeskops P, Viergever MA, Mendrik AM, de Vries LS, Benders MJNL, Isgum I. Automatic segmentation of MR brain images with a convolutional neural network. *IEEE Trans Med Imaging*. 2016;35(5):1252–61.
62. Nalepa J, Marcinkiewicz M, Kawulok M. Data augmentation for brain-tumor segmentation: a review. *Front Comput Neurosci*. 2019a;13:83.
63. Nalepa J, Mrukwa G, Piechaczek S, Lorenzo PR, Marcinkiewicz M, Bobek-Billewicz B, Wawrzyniak P, Ulrych P, Szymanek J, Cwiek M, Dudzik W, Kawulok M, Hayball MP. Data augmentation via image registration. In: Proceedings of IEEE ICIP; 2019b. p. 4250–4.
64. Nalepa J, Myller M, Kawulok M. Transfer learning for segmenting dimensionally reduced hyperspectral images. *IEEE Geosci Remote Sens Lett*. 2020a;17(7):1228–32.
65. Nalepa J, Ribalta Lorenzo P, Marcinkiewicz M, Bobek-Billewicz B, Wawrzyniak P, Walczak M, Kawulok M, Dudzik W, Kotowski K, Burda I, Machura B, Mrukwa G, Ulrych P, Hayball MP. Fully-automated deep learning-powered system for DCE-MRI analysis of brain tumors. *Artif Intell Med*. 2020b;102:101769.
66. Park MTM, Pipitone J, Baer LH, Winterburn JL, Shah Y, Chavez S, Schira MM, Lobaugh NJ, Lerch JP, Voineskos AN, Chakravarty MM. Derivation of high-resolution MRI atlases of the human cerebellum at 3T and segmentation using multiple automatically generated templates. *NeuroImage*. 2014;95:217–31.
67. Park JE, Park SY, Kim HJ, Kim HS. Reproducibility and generalizability in radiomics modeling: possible strategies in radiologic and statistical perspectives. *Korean J Radiol*. 2019;20(7):1124–37.
68. Pesapane F, Codari M, Sardanelli F. Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. *Eur Radiol Exp*. 2018;2(1):35.
69. Pinto A, Pereira S, Correia H, Oliveira J, Rasteiro DMLD, Silva CA. Brain tumour segmentation based on extremely randomized forest with high-level

- features. In: Proceedings of IEEE EMBC; 2015. p. 3037–40.
70. Pipitone J, Park MTM, Winterburn J, Lett TA, Lerch JP, Pruessner JC, Lepage M, Voineskos AN, Chakravarty MM. Multi-atlas segmentation of the whole hippocampus and subfields using multiple automatically generated templates. *NeuroImage*. 2014;101:494–512.
71. Potvin O, Khademi A, Chouinard I, Farokhian F, Dieumegarde L, Leppert I, Hoge R, Rajah MN, Bellec P, Duchesne S, et al. Measurement variability following MRI system upgrade. *Front Neurol*. 2019;10:726.
72. Prastawa M, Bullitt E, Gerig G. Synthetic ground truth for validation of brain tumor MRI segmentation. In: Duncan JS, Gerig G, editors. Medical image computing and computer-assisted intervention – MICCAI 2005. Berlin/Heidelberg: Springer Berlin Heidelberg; 2005. p. 26–33.
73. Ranjith G, Parvathy R, Vikas V, Chandrasekharan K, Nair S. Machine learning methods for the classification of gliomas: initial results using features extracted from MR spectroscopy. *Neuroradiol J*. 2015;28(2):106–11.
74. Renard F, Guedria S, Palma ND, Vuillerme N. Variability and reproducibility in deep learning for medical image segmentation. *Sci Rep*. 2020;10(1):13724.
75. Sahouni M, Kallel F, Dammak M, Mhiri C, Ben Mahfoudh K, Ben Hamida A. A comparative study of MRI contrast enhancement techniques based on traditional gamma correction and adaptive gamma correction: case of multiple sclerosis pathology. In: Proceedings of IEEE ATSIP; 2018. p. 1–7.
76. Saman S, Jamjala Narayanan S. Survey on brain tumor segmentation and feature extraction of MR images. *Int J Multimedia Inf Retr*. 2019;8(2):79–99.
77. Segato A, Marzullo A, Calimeri F, De Momi E. Artificial intelligence for brain diseases: a systematic review. *APL Bioeng*. 2020;4(4):041503.
78. Sheller MJ, Reina GA, Edwards B, Martin J, Bakas S. Multi-institutional deep learning modeling without sharing patient data: a feasibility study on brain tumor segmentation. In: Crimi A, Bakas S, Kuijf H, Keyvan F, Reyes M, van Walsum T, editors. *Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries*. Cham: Springer International Publishing; 2019. p. 92–104.
79. Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, Milchenko M, Xu W, Marcus D, Colen RR, Bakas S. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci Rep*. 2020;10(1):12598.
80. Shen Y, Gao M. Brain tumor segmentation on MRI with missing modalities. *CoRR* abs/1904.07290. 2019.
81. Soltaninejad M, Yang G, Lambrou T, Allinson N, Jones TL, Barrick TR, Howe FA, Ye X. Automated brain tumour detection and segmentation using superpixel-based extremely randomized trees in FLAIR MRI. *Int J Comput Assist Radiol Surg*. 2017;12(2):183–203.
82. Steenwijk MD, Pouwels PJ, Daams M, van Dalen JW, Caan MW, Richard E, Barkhof F, Vrenken H. Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs). *NeuroImage: Clin*. 2013;3:462–9.
83. Stonnington CM, Tan G, Klppel S, Chu C, Draganski B, Jack CR, Chen K, Ashburner J, Frackowiak RS. Interpreting scan data acquired from multiple scanners: a study with Alzheimer's disease. *NeuroImage*. 2008;39(3):1180–5.
84. Sultan HH, Salem NM, Al-Atabany W. Multi-classification of brain tumor images using deep neural network. *IEEE Access*. 2019;7:69215–25.
85. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging*. 2015;15(1):29.
86. Taherdangkoo M, Bagheri MH, Yazdi M, Andriole KP. An effective method for segmentation of MR brain images using the ant colony optimization algorithm. *J Digit Imaging*. 2013;26(6):1116–23.
87. Tarasiewicz T, Nalepa J, Kawulok M. Skinny: a light-weight U-Net for skin detection and segmentation. In: Proceedings of IEEE ICIP; 2020. p. 2386–90.
88. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, Verweij J, Van Glabbeke M, van Oosterom AT, Christian MC, Gwyther SG. New guidelines to evaluate the response to treatment in solid tumors. *JNCI: J Natl Cancer Inst*. 2000;92(3):205–16.
89. Villanueva-Meyer JE, Mabray MC, Cha S. Current clinical brain tumor imaging. *Neurosurgery*. 2017;81(3):397–415.
90. Visser M, Miller D, van Duijn R, Smits M, Verburg N, Hendriks E, Nabuurs R, Bot J, Eijgelaar R, Witte M, van Herk M, Barkhof F, de Witt Hamer P, de Munck J. Inter-rater agreement in glioma segmentations on longitudinal MRI. *NeuroImage: Clin*. 2019;22: 101727.
91. Vreeland A, Persons KR, Primo HR, Bishop M, Garriott KM, Doyle MK, Silver E, Brown DM, Bashall C. Considerations for exchanging and sharing medical images for improved collaboration and patient care: HIMSS-SIIM collaborative white paper. *J Digit Imaging*. 2016;29(5):547–58.
92. Wadhwa A, Bhardwaj A, Verma VS. A review on brain tumor segmentation of MRI images. *Magn Reson Imaging*. 2019;61:247–59.
93. Waghmare VK, Kolekar MH. Brain tumor classification using deep learning. Singapore: Springer Singapore; 2021. p. 155–75.
94. Wen PY, Macdonald DR, Reardon DA, Cloughesy TF, Sorensen AG, Galanis E, DeGroot J, Wick W, Gilbert MR, Lassman AB, Tsien C, Mikkelsen T, Wong ET, Chamberlain MC, Stupp R, Lamborn KR, Vogelbaum MA, van den Bent MJ, Chang SM. Updated response assessment criteria for high-grade gliomas: response assessment in neuro-

- oncology working group. *J Clin Oncol.* 2010;28(11): 1963–72.
95. Wu W, Chen AYC, Zhao L, Corso JJ. Brain tumor detection and segmentation in a CRF framework with pixel-pairwise affinity and superpixel-level features. *Int J Comput Assist Radiol Surg.* 2014;9(2):241–53.
96. Zeineldin RA, Karar ME, et al. Deepseg: deep neural network framework for automatic brain tumor segmentation using magnetic resonance FLAIR images. *Int J Comput Assist Radiol Surg.* 2020;15(6):909–20.
97. Zhao X, Wu Y, Song G, Li Z, Zhang Y, Fan Y. A deep learning model integrating FCNNs and CRFs for brain tumor segmentation. *CoRR* abs/1702.04528. 2017.
98. Zhou T, Canu S, Vera P, Ruan S. Brain tumor segmentation with missing modalities via latent multi-source correlation representation. In: Martel AL, Abolmaesumi P, Stoyanov D, Mateus D, Zuluaga MA, Zhou SK, Racoceanu D, Joskowicz L, editors. *Medical image computing and computer assisted intervention – MICCAI 2020 – 23rd international conference, Lima, Peru, October 4–8, 2020, proceedings, part IV, Lecture notes in computer science, vol. 12264*. Springer; 2020. p. 533–41.
99. Zhu J, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR* abs/1703.10593. 2017.
100. Zhuge Y, Krauze AV, Ning H, Cheng JY, Arora BC, Camphausen K, Miller RW. Brain tumor segmentation using holistically nested neural networks in MRI images. *Med Phys.* 2017;44:5234–43.
101. Zikic D, Glocker B, Konukoglu E, Criminisi A, Demiralp C, Shotton J, Thomas OM, Das T, Jena R, Price SJ. Decision forests for tissue-specific segmentation of high-grade gliomas in multi-channel MR. In: *Proceedings of MICCAI*. Springer; 2012. p. 369–76.
102. Zou KH, Warfield SK, Bharatha A, Tempany CMC, Kaus MR, Haker SJ, Wells William MR, Jolesz FA, Kikinis R. Statistical validation of image segmentation quality based on a spatial overlap index. *Acad Radiol.* 2004;11(2):178–89.



Artificial Intelligence in Stroke

124

Nishant K. Mishra and David S. Liebeskind

Contents

Introduction	1734
Artificial Intelligence	1734
Early Detection of Stroke Symptoms	1736
Acute Stroke Therapy	1737
Role of AI in the Subacute Phase and Follow-Ups of the Ischemic Stroke Patients	1743
Use of AI in the Management of Transient Ischemic Attack	1745
Role of AI in the Management of Intracerebral Hemorrhage	1745
Future Directions	1745
References	1746

Abstract

Artificial Intelligence (AI) is a broad term used to describe the science and engineering of making intelligent machines. AI tools use a range of computer algorithms that learn from experience and allow machines to simulate human intelligence. Management of the stroke patient involves an algorithmic approach. The algorithmic nature of stroke workflow creates an

opportunity for the artificial intelligence (AI)-based tools to increase workflow efficiency. These tools have been implemented in the stroke care and research and include tools for the patient triage, image analysis, decision making with therapeutics and interventions, rehabilitation, and trial design. In this chapter, we describe the critical aspects of the stroke management, clinical contexts where the AI tools have been used, and suggest scientific questions that should be tackled using AI.

N. K. Mishra

Department of Neurology, UCLA Stroke Center,
University of California, Los Angeles, CA, USA

D. S. Liebeskind (✉)

Vascular Neurology, University of California, Los
Angeles, CA, USA

Keywords

Stroke management · Perfusion Imaging ·
Thrombolysis · Endovascular recanalization ·
Mismatch · Machine learning

Introduction

Stroke is the fifth major cause of mortality in the United States [1]. It is broadly classified into ischemic stroke and hemorrhagic stroke. Ischemic stroke accounts for about 80% of the stroke patients. It is caused by the interruption of blood supply to the brain. Hemorrhagic stroke is caused by the rupture of a blood vessel or of an abnormal vascular structure. The differentiation into ischemic and hemorrhagic stroke is the first critical step in the management of ischemic stroke patients because the treatment approach is different for these two stroke subtypes. For the ischemic stroke patients, the goal is to recanalize the brain vessel feeding into the ischemic territory and thus achieve rapid reperfusion of the hypo perfused but salvageable brain parenchyma. Eligible patients are, therefore, offered intravenous (iv) thrombolytic agents and/or recanalization therapy [2]. In case of intracerebral hemorrhages, the use of iv thrombolytic agents is contraindicated, and the goal is to prevent clinical worsening due to mass effect from the hematoma (Fig. 1) [3].

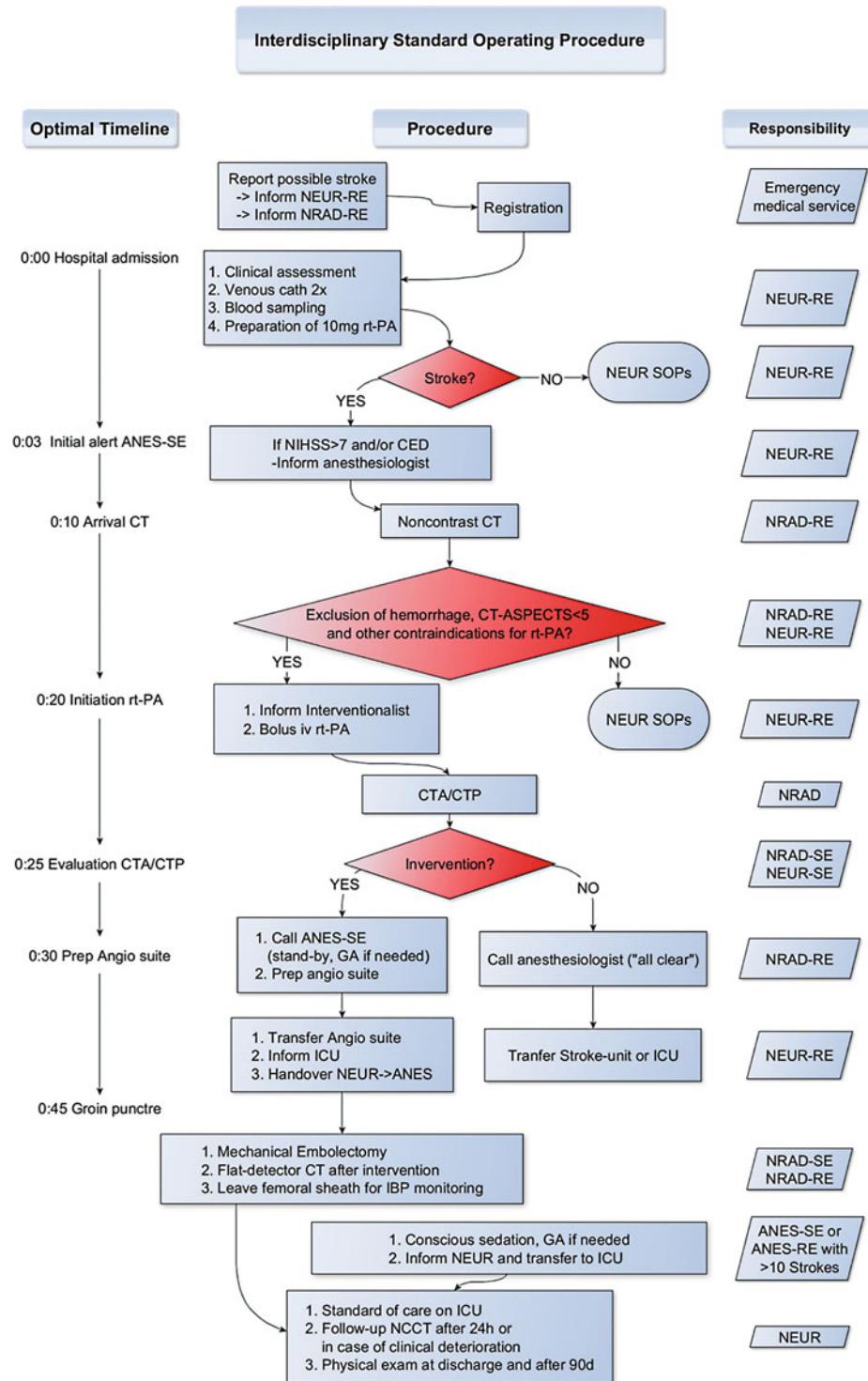
Intravenous alteplase is administered to eligible ischemic stroke patients who present within the prescribed time window (i.e., within 4.5 h of the symptom onset), do not have contraindications to iv alteplase, and do not show intracranial bleed on immediate brain imaging. Two NINDS trials demonstrated the safety and efficacy of iv alteplase in 3 h time window in acute ischemic stroke patients and were reported together in 1995 [5]. A pooled analysis subsequently showed that the odds ratio for the improved outcome declines rapidly after the symptom onset (Fig. 2) [6]. The need to quickly offer iv alteplase to eligible ischemic stroke patients led to the recognition that it was critical to avoid delays. With that objective, Hazinski et al. proposed a “stroke chain of survival and recovery” in 1997 [7]. The stroke chain of survival and recovery involves *seven Ds*: **d**etection (prompt detection of stroke symptoms), **d**ispatch (quick dispatch of the emergency services to a patient’s location), **d**elivery (rapid administration of prehospital care and quick transport of the

patient to an appropriate stroke center), **d**oor (efficient triage of the patient at the emergency department of the stroke center), **d**ata (quick clinical assessment and neuroimaging), **d**ecision (rapid interpretation of the clinical and radiological data and determination regarding which treatment pathway to pursue), and **d**rug therapy (decision to offer reperfusion therapy, or other therapy) [7, 8]. Rehabilitation is the next step in the stroke chain of recovery wherein the goal is to optimize a patient’s functional outcomes and integrate them into the social milieu. The algorithmic nature of stroke workflow creates an opportunity for the artificial intelligence (AI)-based tools to increase workflow efficiency.

AI-based tools have been implemented in the stroke care and research and include tools for the patient triage, image analysis, decision making with therapeutics and interventions, rehabilitation, and trial design. In this chapter, we describe the critical aspects of the stroke management, clinical contexts where the AI tools have been used, and suggest scientific questions that should be tackled using AI tools.

Artificial Intelligence

AI is a broad term used to describe the science and engineering of making intelligent machines. AI tools use a range of computer algorithms that learn from experience and allow machines to simulate human intelligence (chapters that inform technical aspects of AI). Deep learning is a subset of machine learning which in turn is a subset of AI. Machine learning-based imaging tools demonstrate their ability to rapidly perform objective and quantitative inspection of the subtle differences within and between the voxels, and these can be repeatedly performed on a large scale. Artificial Neural Networking (ANN) is an example of supervised machine learning that is inspired by the biological neural units that process the data in a manner similar to the brain neurons, that is, by the activation and inhibition of neurons. It can be used to build prediction models, and its advantage

**Fig. 1** An example of stroke care workflow [4]

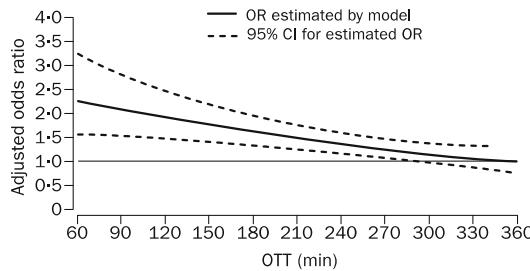


Fig. 2 Pooled individual patient data analysis of the NINDS, ECASS, and ATLANTIS trials showing adjusted odds ratio for favorable outcome at 3 months in ischemic stroke patients treated with iv alteplase [9]

is that it requires less input, less formal statistical training, and can handle nonlinear relationships between the variables.

Early Detection of Stroke Symptoms

The years 1995 and 2008 are the turning point in the history of stroke care when the NINDS trials and the ECASS III trial reported data to support the use of iv alteplase in eligible ischemic stroke patients up to 4.5 h of the symptom onset [5, 10]. Since 2015, several randomized trials have also shown that the patients with large vessel occlusion also benefit from mechanical thrombectomy up to 18 to 24 h of the symptom onset [11–17]. These data have led to major restructuring of the workflow to offer stroke care. The goal is to offer quick delivery of iv thrombolytic agent and/or endovascular recanalization to an eligible patient [18–21].

In the prehospital stage, a paramedic or mobile stroke unit team identifies the stroke symptoms, performs early assessment and information gathering, and in case of mobile stroke unit even administers iv alteplase, and then quickly prompts an appropriate stroke center – primary or comprehensive – with relevant clinical information [22–24]. Comprehensive stroke centers are chosen if the patient is potentially a candidate for endovascular treatment [25–28]. The communication is one way in this prehospital triage, and it is critical that the information sharing is not fragmented, recurrent, or disorganized [21, 29, 30]. The stroke care heavily relies on synchronized flow of information

throughout the stroke care pathway, and delays can occur at various levels during the stroke care pathway (Table 1) [21]. Automated communication system during the prehospital management can better prepare the receiving hospital to allocate resources as soon as the patient enters their facility [21]. An Australian group has reported a python-based electronic communication system to streamline code stroke alert pathway. This communication system has data fields relevant to each stage of the chain of stroke care to which data relevant to a provider's role can be input [21]. At the receiving end, this platform provides a tiered level access to data based on the role of the individual in the stroke care pathway [21]. Within this automated platform, AI-based tools can be integrated at each step to assist with decision making.

When called to evaluate a patient with stroke like symptom(s), the first major step is to ascertain that the patient's symptoms are truly of vascular origin. Missed diagnosis of acute stroke accounts for about 13% of the stroke presentation to an emergency department (ED) [31]. An analysis of ED data revealed that the patients with intracranial hemorrhage who presented with headache and ischemic stroke patients who presented with dizziness and headache were more likely to be missed [31]. Abedi et al. [32] built an artificial neural network (ANN) model that correctly differentiated acute cerebral ischemia from stroke mimic with a 92% precision. They used Neuralnet package of the freely available R software for the model development and applied back propagation-based learning followed by tenfold cross validation [32]. We propose that the ANN models like this can be integrated with in the electronic medical records of hospital networks to quickly analyze patient's data and advise the stroke team about the level of confidence in predicting a vascular cause of the patient's symptoms.

Patients with large vessel occlusion are offered endovascular recanalization and it is desirable to be able to select a patient population with greater likelihood of the presence of large vessel occlusion predominantly based on the clinical data available at prehospital stage. Chen et al. [33] reported an artificial neural network (ANN) algorithm that was associated with a high accuracy

Table 1 Common causes of the delay in reperfusion therapy [21]

Community	EMS	Hospital
Barriers to Timely Reperfusion		
Delayed or mistaken identification of stroke	Delay in arrival or inappropriate triage	Delay in initial assessment or triage
Delay in alerting EMS	Delayed assessment and information gathering	Delay in imaging acquisition or review
	Delayed decision-making	Delayed decision making
	Transport to hospital	Delay in activation of mechanical thrombectomy team
		Delay in transfer from the emergency department to angiography
		Delay in transfer from an ED to another mechanical thrombectomy-capable hospital

(82%), sensitivity (83%), and specificity (82%) for predicting large vessel occlusion (LVO) in the prehospital stage (see Table 2). This algorithm used the data available at the prehospital stage and included a patient's demography, National Institutes of Health Stroke Scale (NIHSS), and the patient's vascular risk factors [33]. The authors included patients who received reperfusion therapy within 8 h from symptom onset and assessed their CT or MR angiogram for the presence of occlusion in the internal carotid occlusion, M1 and M2 branch of the middle cerebral artery, and basilar artery [33]. This was a retrospective analysis and the stroke patients in this study had received reperfusion therapy. A better approach would be to design an AI-based model that uses large data of unselected stroke patients in the prehospital stage and prospectively determines the hidden predictive features to optimally classify patients into LVO and no LVO group.

Acute Stroke Therapy

When a patient with stroke symptoms arrives at a stroke center, patient's cardiorespiratory stability is ascertained and brain imaging is immediately performed; the goal is to determine whether the patient had a hemorrhagic stroke [34–37]. If there is no evidence of intracranial hemorrhage, the patient is a candidate for reperfusion therapy provided he has no additional contraindications [38, 39]. There is a potential to automate the steps leading to the administration of

thrombolytic agents and facilitate the treating physician. The automation would involve automated retrieval of patient's medical history and medication refill history to supplement the information obtained during the stroke code.

NINDS trials and ECASS III trials supported the use of iv alteplase within 3 h and 3–4.5 h of the symptom onset. Time since symptom onset is uncertain in patients with wakeup stroke or unwitnessed stroke. Mismatch between the Diffusion weighted imaging signal and fluid attenuated inversion recovery imaging signal (DWI-FLAIR mismatch paradigm) was proposed to select patients when the time since symptom was uncertain [40]. DWI-FLAIR mismatch is present if the patient's magnetic resonance imaging (MRI) reveals hyperintensity on DWI sequence and no hyperintensity on the FLAIR sequence and indicates that the patient is within the time window for receiving thrombolytic agent [41]. This approach, however, has limitation: this patient selection approach has limited sensitivity (0.78) and specificity (0.62); and interobserver agreement (0.57) is moderate [40]. Because of the limited sensitivity and moderate interrater agreement, DWI-FLAIR mismatch paradigm is not reliable. Deep learning approaches can, however, be applied to learn the “deep” features on DWI and FLAIR sequences; this approach can be used to correctly classify the patients for the patient's time since last seen normal and improve the reliability of DWI-FLAIR mismatch paradigm. Ho et al. [42, 43] used auto-encoder architecture for deep learning and demonstrated that stepwise multilinear regression

Table 2 Diagnostic parameters of the prehospital prediction scales compared to the Chen's artificial neural network model [33]

	FAST-ED	3-ISS	RACE	PASS	CPSSS	LAMS	NIHSS	NIHSS ≥ 6	ANN	ANN ^a
AUC	0.783	0.782	0.776	0.784	0.796	0.740	0.790	/	0.823 \pm 0.060	0.804 \pm 0.042
Youden index	0.467	0.453	0.427	0.493	0.490	0.403	0.453	0.327	0.640 \pm 0.105	0.557 \pm 0.067
Sensitivity	0.760	0.583	0.730	0.727	0.713	0.807	0.607	0.847	0.807 \pm 0.071	0.729 \pm 0.081
Specificity	0.707	0.870	0.697	0.767	0.777	0.597	0.847	0.480	0.833 \pm 0.060	0.828 \pm 0.106
Accuracy	0.733	0.727	0.713	0.747	0.745	0.702	0.727	0.663	0.820 \pm 0.053	0.778 \pm 0.033
Cutoff	3	3	4	2	2	3	3	12	/	/

The ANN model included age, gender, prior antiplatelet therapy, 15 NIHSS items and nine risk factors, while ANN model^a only included 15 NIHSS items. FAST-ED indicates Field Assessment Stroke for Emergency Destination, 3I-SS indicates 3-item Stroke Scale, CPSSS indicates Cincinnati Prehospital Stroke Severity scale and LAMS indicates Los Angeles Motor Scale, PASS indicates Prehospital Acute Stroke Severity scale.

(SMR), support vector machine (SVM), random forest (RF), and gradient boosted regression tree (GBRT) were able to improve the classification of the stroke patients into a iv alteplase eligible group (time since symptom onset, i.e., <4.5 h) and ineligible group (time of symptom onset $>=4.5$ h) [42]. To be able to determine patient's time since symptom onset based on AI algorithm, and not merely on the family and EMS, will hasten the stroke work flow and also augment a physician's confidence on this key variable which is used to determine whether to administer a thrombolytic agent [42, 43].

Pooled analyses of iv alteplase trials revealed that the odds for improved functional outcomes decline rapidly over time, but stay statistically significant in the first 4.5 h of the symptom onset (Fig. 1) [9, 44]. In the time window 4.5 h to 6 h, the odds for improved outcomes are statistically nonsignificant and associated with a wider confidence interval (CI), that is, 0.9 to 1.5 [6, 9, 44, 45]. (Fig. 1) [44]. This suggests that in the time window beyond 4.5 h, there may still be patients who could benefit from thrombolysis; others may, however, be at an increased risk from delayed reperfusion therapy [46]. In order to select patients in delayed time window, a perfusion and ischemic core mismatch paradigm (PWI/DWI mismatch, for example) was proposed with a goal to identify ischemic stroke patients with persistent salvageable tissue (penumbra) [46, 47]. This model was supported by earlier EEG and PET studies. For example, in humans, the clamping of carotid artery results in flattening of the electroencephalogram (EEG) at the hemispheric blood flow below 0.16–0.17 ml/g/minute [48–50]. Cerebral blood flow below 0.10 ml/g/minute is associated with infarction (ischemic core), cerebral blood flow in the range 0.10 to 0.17 ml/g/minute is associated with ischemic penumbra (hypo perfused but salvageable brain tissue), and cerebral blood flow greater than 0.17 ml/g/minute is associated with the presence of benign oligemia or normal tissue [48–56]. The goal of reperfusion therapy is to prevent the growth of ischemic core into the penumbra. The goal of mismatch paradigm is to be able to distinguish the benign oligemia from true “penumbra.” An ideal approach to make this distinction would be to use

positron emission tomography (PET) imaging. PET can be used to quantify the CBF, regional oxygen extraction fraction, and the regional metabolic rate for oxygen. PET, however, is cumbersome in routine clinical setting; alternative approaches like MR- or CT-based perfusion imaging were therefore proposed. MR and CT perfusion imaging provide the perfusion metrics like time to peak (TTP), mean transit time (MTT), Time to peak (Tmax), cerebral blood volume (CBV), and CBF. These are calculated using the concentration time curve derived from a patient's cerebral artery and/or vein (Fig. 3) [57, 58].

A meta-analysis of the DIAS, DIAS-2, DEDAS, EPITHET, and DEFUSE trials revealed that mismatch-based delayed treatment resulted in the greater odds for recanalization; this, however, did not translate into improved clinical outcome with thrombolysis in 3–6 h time-window [46]. Some of the challenges with the earlier mismatch paradigm were the lack of automation during the postprocessing of the perfusion images (e.g., manual selection of the arterial input function), conservative thresholds of Tmax and mismatch ratio that therefore also included regions with benign oligemia, assumption of linear relationship between the signal change and concentration time curve, and eyeballing of lesion volume [46, 59–62]. Several challenges related to the use of the postprocessing of the perfusion imaging data have now been tackled. For example, contemporary postprocessing soft wares are now able to automatically correct the perfusion images for the errors due to motion and time of each imaging sequence, automatically select the location for measuring concentration time curve, correct for the nonlinear relationships related to the gadolinium in case of MR perfusion, perform deconvolution, compute perfusion maps for coregistration with the ischemic core maps (DWI in case of MR paradigm and rCBF $<30\%$ in case of CT perfusion), and provide mismatch summary maps (Fig. 4). Postprocessing softwares are now available through various vendors like IschemiaView (RAPID), Olea, and VizAI.

Automated postprocessing of perfusion imaging data has been a major advance in the use of

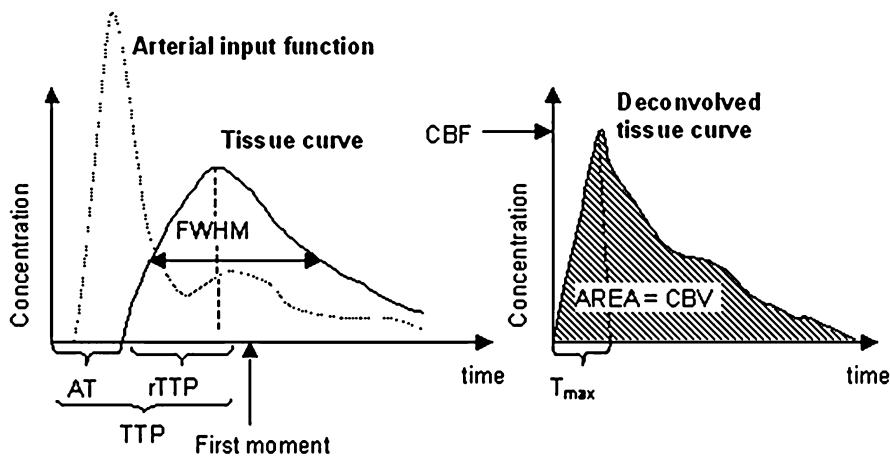


Fig. 3 Concentration time curve (time to peak (TTP), peak time fitted minus arrival time fitted; AT: bolus arrival time; FWHM (full width half maximum): width of the

concentration time curve at the point of halfway to the maximum concentration; 1st Moment: balancing point of the curve; AUC: Area under curve) [57, 58]

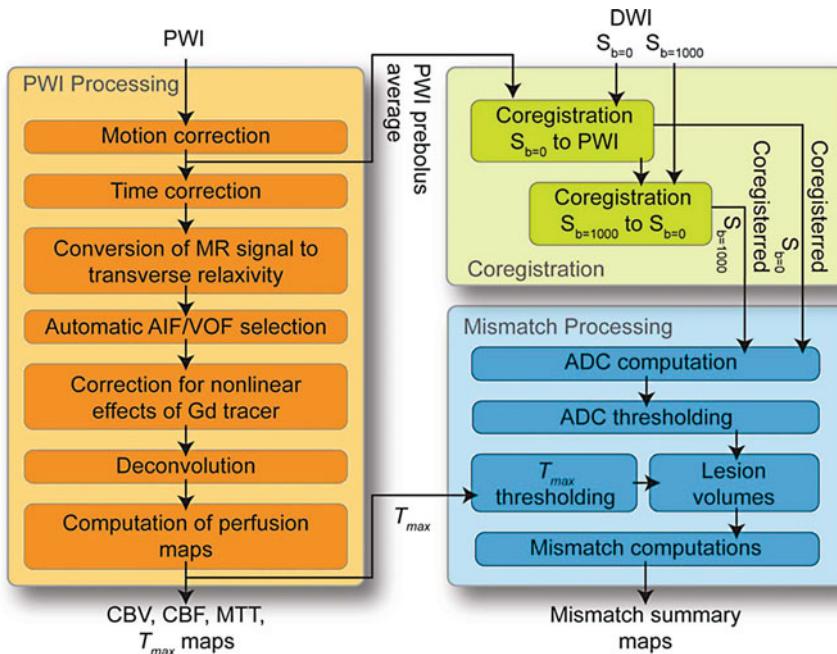


Fig. 4 Steps involved in the postprocessing of MR perfusion data by Stanford's RAPID software [63]

mismatch paradigm to select patients for reperfusion therapy. Automated detection of the concentration time curves, that is, the arterial input function (AIF) and venous output function (VOF), is an example of the use of artificial intelligence in perfusion imaging [63]. The automated, operator independent, selection of concentration

time curves (AIF and VOF) has removed the interrater variability which used to arise from manual selection and thus has given consistency to the calculation of the perfusion metrics [63]. Automation also saves the time in the post-processing of the perfusion data as this was a major challenge that precluded the use of

mismatch paradigm in selecting patients in the trials like EPITHET [46]. Matus Straka and colleagues at Stanford developed a postprocessing software called RAPID which has been used in several perfusion imaging trials. RAPID's algorithm automatically selects an optimal AIF and VOF locations by the following steps: first, it automatically identifies the height, arrival time, and the width of the concentration time signal at several brain locations and determines their average values; then the perfusion algorithm searches for the spatially clustered locations where the concentration time signals have an above average amplitude, below average width, and early bolus arrival time; and finally, the algorithm determines the best fit for these clustered concentration-time curves using the weights determined from the perfusion imaging data [63]. The algorithm ensures that the input function estimates are optimized for anatomically appropriate locations, that is, the large arteries and veins [63]. Thus, the RAPID software determines the location of the final input function over a vessel where the sum of the fit statistic (c) is maximum, and this includes a location with at least three adjacent pixels with the dimensions $1.88 \text{ mm} \times 1.88 \text{ mm}$ and six adjacent pixels with dimensions $0.94 \text{ mm} \times 0.94 \text{ mm}$ [63].

Despite being a major advance in the field of mismatch-based reperfusion therapy, the perfusion imaging algorithms are not without challenges. For example, whereas we use common perfusion thresholds for gray and white matter, there are differences between the two [64]. Relative CBF $<30\%$ best defines core in the gray matter (AUC 0.73) and $<20\%$ best defines the core in the white matter (AUC 0.67) [64]. Similarly, a Tmax delay of $>5 \text{ s}$ best defines a perfusion lesion in gray matter (AUC 0.80) and a delay of $>7 \text{ s}$ best defines a perfusion lesion in white matter (AUC 0.75) [64]. Postprocessing therefore ought to automate methods to select optimal thresholds for gray and white matter and determine core and penumbra threshold specific to the location of the individual brain region being analyzed.

Another challenge is that the postprocessing of the imaging data involves assumptions and

thresholds which are prone to introduce errors. Assumptions are involved in deconvolution of the AIF, the concentration time curve from artery, to generate residue function; and residue function is then used to calculate the perfusion metrics by curve fitting approach, a process which is prone to errors [65–67]. A major challenge is to calculate AIF in a voxel from the small vessel feeding into each voxel; a global AIF from a large vessel like a middle cerebral artery is therefore used. Independent component analysis has been reported to demonstrate how the local AIFs can be selected for postprocessing [68]. The characteristics of the AIF curve depend upon the patient-specific cerebrovascular anatomy and the cardiac function [66]. The variation in the local AIFs and cerebrovascular anatomy, physiologic changes from, for example, cardiac dysfunction, and the use of thresholded perfusion metrics like Tmax and relative CBF contribute to the errors in the measurement of perfusion metrics [69].

A major challenge with the contemporary perfusion imaging softwares is the significant reliance on the use of fixed thresholds of the perfusion metrics (e.g., $\text{Tmax} > 6 \text{ s}$ to define penumbra and $\text{rCBF} < 30\%$ to define core on CT perfusion). The deep convolution neural network-based algorithm can avoid the need for using thresholds by determining the tissue fate by learning the features directly from the concentration time curve thus not requiring the AIFs and the assumption-based curve fitting [69]. For example, Ho et al. applied a convolutional neural network-based AI approach on 4 dimensional (x, y, z, t) perfusion imaging data to predict final infarct volume [69]. They were able to show significantly improved prediction of tissue fate [69]. Deep convolutional neural networking has major strength in its ability to learn the data driven filters to extract complex features linked to tissue infarction. This approach is better than using model derived perfusion metrics like CBF because it does not require the use of error prone AIF or the assumptions used in deconvolution [69].

Regarding the choice of thresholds, a mismatch ratio >1.8 and Tmax delay threshold of $>6 \text{ s}$ is in use to define mismatch profile [70, 71]. These thresholds are not fixed, however.

The earlier perfusion imaging studies used to use a mismatch ratio of 1.2 and Tmax delay threshold of >2 s to define mismatch [46]. The older trials that used these conservative thresholds were found to be negative, and a common element noted for the failure of these trials was that these thresholds included patients with benign oligemia [46]. The variation in the thresholds of the perfusion metrics is due to the significant heterogeneity between the patient cohorts, variations in the postprocessing of the imaging data, and simplistic assumptions regarding the tissue physiology (e.g., the presence of collaterals, tissue type, lesion location, vascular delays, and dispersion effects). Because of the complex nature of these perfusion metric, certain thresholds are sometimes chosen out of trial design considerations. A secondary analysis of the DEFUSE [72] data showed that a mismatch ratio of 2.6 is associated with maximum sensitivity; [71] the DEFUSE trialists, however, selected a mismatch ratio of 1.8 to define the ischemic penumbra in the subsequent DEFUSE 2 [73], CRISP [74], and DEFUSE 3 [17] studies, out of the trial enrolment consideration [17, 73, 74]. The problem of a fixed threshold is greater with CT-based perfusion imaging paradigm where not only the perfusion metric is derived from thresholding, the ischemic core is also determined from an optimized threshold of relative CBF, and a threshold of $<30\%$ is commonly considered to be an optimal value [75]. In the sphere of perfusion imaging, DWI hyperintensity is considered a gold standard to define ischemic core; however, this metric is also not without challenges. For example, diffusion reversal, even though is rare, does occur [76–80]. Automated segmentation of DWI lesions is prone to errors as well [77, 78]. For example, the $b = 1000$ images are preferred by human eyes to visualize the DWI lesions, but these cannot be used to automate segmentation using computing software [63]. RAPID, for example, uses an ADC threshold of $<600 \times 10^{-6} \text{ mm}^2/\text{s}$ to automatically define DWI core arguing that this threshold is more acceptable for computerized segmentation and removes the noise due to the artifacts like T2 shine through, susceptibility pile up, and coil sensitivity field variation [63]. In order to avoid the need for

using fixed thresholds, convolution neural network can be used to yield more reliable volumes and location of ischemic core. Chen et al. [81] have attempted this AI approach: they used the first fully automated deep convolutional neural networks to segment the acute ischemic stroke lesions on DWI. They then validated their approach on a large clinical dataset achieving good dice coefficient (mean: 0.67; small lesion: 0.61; large lesion: 0.83) [81]. Chen et al. also observed that the approach by combining the EDD Net and multiscale convolutional label evaluation (MUSCLE) Net approach gave the best results [81].

Perfusion imaging technology not only can identify mismatch; it can also prognosticate the risk of the ischemic brain regions to develop hemorrhagic transformation, the most feared complication of reperfusion therapy. For example, the brain regions with significant Tmax delays, blood brain barrier disruption, and very low cerebral blood volumes have been shown to be at an increased risk of hemorrhagic transformation and poor outcomes [82–85]. AI tools can be used to learn the deep features of the Tmax delays, low CBV, and permeability changes to determine the risk of hemorrhagic transformation. AI tools should also be designed to provide an easy to interpret metric that informs a physician regarding its “confidence” on the output that the algorithm generates.

It can be argued that the perfusion-diffusion mismatch paradigm is a simplistic dichotomous predictive model of patient outcome [71]. Mismatch ratio has both interindividual and intervoxel variability; a common threshold should therefore not be applied. PET-based selection of mismatch would be more reliable, but it is challenging to use it in an acute stroke setting [57, 58]. It is, however, possible to train the deep learning algorithms like deep convolutional neural network on patients with subacute and chronic ischemia and then apply this model in acute stroke setting for defining optimal volume of mismatch.

MR CLEAN proved that the showing of arterial occlusion at stroke onset was critical to the success of endovascular treatment [11]. The site of arterial obstruction strongly predicts patient

outcomes [86]. Endovascular recanalization relies heavily on the prompt detection of the intracranial large vessel occlusion (LVO). Several vendors have developed automated software for this purpose. The presence of an LVO is associated with nonopacification of the vessel distal to the occlusion and results in reduced vessel density on the side of vessel occlusion [87]. Algorithms built for this purpose look out to detect the hemispheric differences in the vessel densities within the suprasellar cisterns and sylvian cistern [87, 88]. The RAPID applies the following steps in the detection of LVO: perform motion and tilt correction of the CTA raw data; align the human head template with the CTA data; create masks by warping the templates off of the anatomic structures like the bones; use bone mask to remove the skull base and calvarium; dichotomize the intracranial vessels into large and small vessels based on the vessel dimensions; calculate the vessel density and compare the vessels to their contralateral counterpart, first with in suprasellar and proximal sylvian cistern, then distally; build maximum intensity projections (MIPs) of the vasculatures from the bone masked CTA; and apply color code to the vessels depending on the difference in the interhemispheric vessel density on the MIP images [87]. To detect an LVO, vessel density threshold of 75% or less is applied by the RAPID software [87]. Amukotuwa et al. [87] reported 92% sensitivity, 97% negative predictive value, and 81% specificity for the detection of intracranial LVO or M2 segment vessel occlusion using the RAPID's fully automated LVO-detection tool. The study highlighted that the false positives commonly resulted from the vascular asymmetry [87]. RAPID's processing time is $</=5$ min [87]. Automated detection of LVO is particularly useful at the smaller centers where radiology services are not always available, and the findings of LVO by this software can automatically alert the nearest comprehensive stroke center initiating a prompt transfer of the patient to a center equipped with tools and expertise to provide mechanical thrombectomy [87].

Because mechanical thrombectomy is now offered as a standard of care, it is now possible

to investigate the pathologic and radiologic characteristics of the cerebral clots causing stroke [89, 90]. A clot perviousness can be calculated by subtracting the CT attenuation measured using noncontrast CT head and CTA head from the clot location [91]. Both the functional outcome and successful thrombectomy are linked to the perviousness of the clot in the cerebral vasculature [91–94]. The clot perviousness can guide an interventionist regarding whether the clot will lyse from the iv or intra-arterial thrombolytic agent or whether mechanical thrombectomy would be needed. AI-based approach can automate the measurement of clot perviousness and incorporate it in the CT/CTA head protocol for use in routine clinical practice.

Successful recanalization is not an equivalent of successful reperfusion. Despite successful recanalization, patients do not show improved outcomes. The clinical outcomes depend not only on the degree of penumbra salvaged or the degree of vessel recanalized; they also depend on the other clinical variables like the oxidative stress from the interaction of hyperglycemia and ischemia or the activation of inflammatory cascades. Large clinical database linked to patient images should be exploited to better automate AI-based prediction of patient outcomes.

Role of AI in the Subacute Phase and Follow-Ups of the Ischemic Stroke Patients

The goal of stroke management in the subacute phase is to determine the stroke mechanism. This is done regardless of whether a stroke patient received reperfusion therapy or not. It is critical to determine the stroke mechanism because it guides the physician to adopt measures to prevent stroke recurrence. A popular classification of the stroke mechanism is the Trial of Org 10172 in Acute Stroke Treatment (TOAST) criteria and includes the following subtypes: (1) large-artery atherosclerosis (embolus or thrombosis); (2) cardio embolism (high-risk or medium-risk); (3) small-vessel occlusion (e.g., lacune); (4) stroke of other determined etiology; (5) stroke of

undetermined etiology (with either two or more causes identified, negative evaluation, or Incomplete evaluation) [95]. Based on the degree of certainty, these subtypes are referred as possible or probable [95].

Early detection of atrial fibrillation in stroke patients is critical because the showing of atrial fibrillation indicates that the stroke mechanism was cardio embolism and the patient should be started on anticoagulation therapy to prevent stroke recurrence linked to cardioembolism. Early identification of atrial fibrillation is also critical because its management can prevent heart failure and death in these patients. Hence, these patients are started on continuous electrocardiogram (ECG) monitoring during their hospital stay, and if no atrial fibrillation is detected by the time of their discharge, a loop recorder is often applied to remotely monitor cardiac rhythm. An AI algorithm can be used to carefully analyze the hidden ECG patterns to detect the risk of atrial fibrillation much earlier than they are detected in routine ECG monitoring. Attia et al. [96] developed an AI-enabled ECG using convolutional neural network to detect the ECG signature of atrial fibrillation. This analysis was based on a very large database: 180,922 patients with 649,931 normal sinus rhythm ECGs [96]. A single AI-enabled ECG identified atrial fibrillation with 79.4% accuracy, 79% sensitivity, and 79.5% specificity [96]. When the ECGs obtained during the first month after the stroke onset were analyzed, the accuracy increased to 83.3%, sensitivity to 82.3%, and specificity to 83.4% [96]. Kashaou et al. [97] reported a case of a 92-year-old woman with hypertension, diabetes mellitus, and peripheral arterial disease patient who had recurrent cryptogenic strokes (left frontal and five years later a posterior circulation stroke with acute right leg ischemia) but repeat ECGs and cardiac monitoring found sinus rhythm. Her echo at the time of first stroke revealed left atrial enlargement, but because of no evidence of atrial fibrillation on ECG or outpatient Holter monitoring, she was continued on antiplatelet therapy [97]. Her second stroke occurred 5 years later [97]. This time, transesophageal echocardiogram revealed left atrial enlargement with left atrial thrombus, and therefore, she was started on anticoagulation

[97]. The ECG and cardiac monitoring at the second stroke again showed sinus rhythm [97]. A retrospective AI-enabled analysis of her prior ECGs revealed that the earliest detection of atrial fibrillation risk could have occurred 12 years prior to her first thromboembolic event [97]. Despite an elevated CHA₂DS₂Vasc score, this patient could not be started on anticoagulation therapy after her first stroke. It is unclear what exactly the AI enabled ECG sees that human eyes do not; however, validation of this approach in other stroke patient population will potentially pave the way for primary stroke prevention in patients likely to develop cardioembolic stroke from an atrial fibrillation.

Whereas stroke mechanisms linked to large artery atherosclerosis or cardio embolism can relatively easily be determined by the means of vessel imaging and telemetry (or loop recorder at discharge), about 25% of the ischemic stroke patients continue to have an undetermined stroke mechanism; this group has historically been referred as having a cryptogenic stroke [98]. Embolic Stroke of Undetermined Source (ESUS) is a recent construct developed to identify and enroll ESUS patients in anticoagulation clinical trials [99]. ESUS is defined as the stroke subtype in which the patients have a nonlacunar brain infarct, vessel imaging reveals <50% luminal stenosis in the arteries supplying the ischemic region, and thorough work-up is unable to reveal a major risk of cardioembolic source or other specific cause of stroke like arteritis, dissection, vasospasm, or drug abuse. Potential embolic sources in ESUS are (a) atrial cardiopathy (defined on echocardiogram as left atrial dilatation or increased left atrial diameter (>38 mm in women and >40 mm in men); or on ECG as the presence of supraventricular extrasystoles), (b) atrial fibrillation, (c) left ventricular (LV) dysfunction (defined as low LV ejection fraction (<35%), LV hypertrophy, left sided heart failure, or LVH on ECG using Sokolow's index >/=35 mm), (d) cardiac valvular disease (moderate to severe stenosis or regurgitation), (e) patent foramen ovale, and (f) cancer [100, 101]. AI-based algorithms can be used to determine stroke mechanism, including the potential sources of embolism in ESUS. For example, Ntaios et al. [102] applied unsupervised machine learning approach, namely,

hierarchical clustering, on the data from three registries (Acute Stroke Registry and Analysis of Lausanne (ASTRAL), Athens Stroke Registry, and Larissa Stroke Registry), to investigate the potential sources of embolism in ESUS patients. They determined four optimal clusters using data-driven machine learning analysis and identified clusters strongly associated with arterial disease, arterial cardiomyopathy PFO, and left ventricular disease [102]. Kamel et al. [103] used machine learning approach on the patients lodged within the Cornell Acute Stroke Academic Registry (CAESER) to distinguish cardioembolic vs. noncardioembolic etiology. This distinction is critical because of the therapeutic implications. Stroke etiology was determined in 1663 patients (large artery atherosclerosis, $N = 291$; cardiac embolism, $N = 688$; other determined etiology, $N = 104$) and cryptogenic in 580 patients [103]. Kamel et al.'s predictive model distinguished cardioembolic from noncardioembolic with a very high validity (AUC 0.85) among the 1663 patients with determined stroke etiology [103]. They used an ensemble machine learning method known as super learner algorithm to subsequently compute a Bayesian targeted machine learning estimator. This machine learning estimator detected that 44% of the ESUS cases were due to occult cardioembolic source [103]. Among the 580 patients who were classified as ESUS patients, the machine learning estimator found a significantly increased predicted probability of cardioembolic source and eventual detection of atrial fibrillation.

Use of AI in the Management of Transient Ischemic Attack

Transient ischemic attack (TIA) is a major risk factor for stroke [104, 105]. It is critical to rapidly identify it and take measures to reduce the risk of subsequent TIAs and/or stroke. Perfusion imaging, including artificial spin labeling, can play an important role in its detection [106]. Prolonged Tmax or MTT, for example, can support that the presenting semiology is neurovascular (i.e., TIA) even though the clinical symptoms were transient. Approximately one fourth of the TIA patients are also found to have a DWI lesion, and this changes

the diagnosis from TIA to silent stroke [107]. Perfusion imaging-based TIA detection algorithm can be incorporated in stroke workflow for AI-guided automated detection of patients with TIA [106, 108–110]. For example, in a study that use the convolutional neural network to predict the cause of TIA like presentation, the AUC was greatest ($88.3 +/− 3.6$) when the network was given the patient's presenting complaint and MRI report [110].

Role of AI in the Management of Intracerebral Hemorrhage

Spontaneous intracerebral hemorrhage (ICH) accounts for about 20% of the strokes and is associated with very high mortality [111]. Despite significant efforts, there has not been major success in therapeutic management of these patients. A management workflow is shown in Fig. 5.

Hematoma volume is a significant predictor of outcomes. About 30% of the cerebral hematoma expand over time and are associated with further clinical worsening because they worsen the mass effect. It is critical to identify patients who are more likely to develop hematoma expansion. AI approach to detect this patient population would be useful. Liu et al. [113] have reported a support vector machine-based AI algorithm which learns from observing the ICH datasets ($N = 1157$ patients) and intrinsically builds complex models to determine relationships between the covariates and the outcome of interest (i.e., hematoma expansion). Automated stratification of a patient's predicted outcome will not only inform the patient's family with regards to the goal of care discussion but also held design better trials.

Future Directions

Prompt treatment of acute stroke, its cause, and complication is critical to reducing the morbidity and mortality. Futures studies that validate the AI approach are needed to allow for its greater acceptance in the routine clinical practice. The goals of AI in stroke management are to facilitate the physicians and not replace them, and it is expected

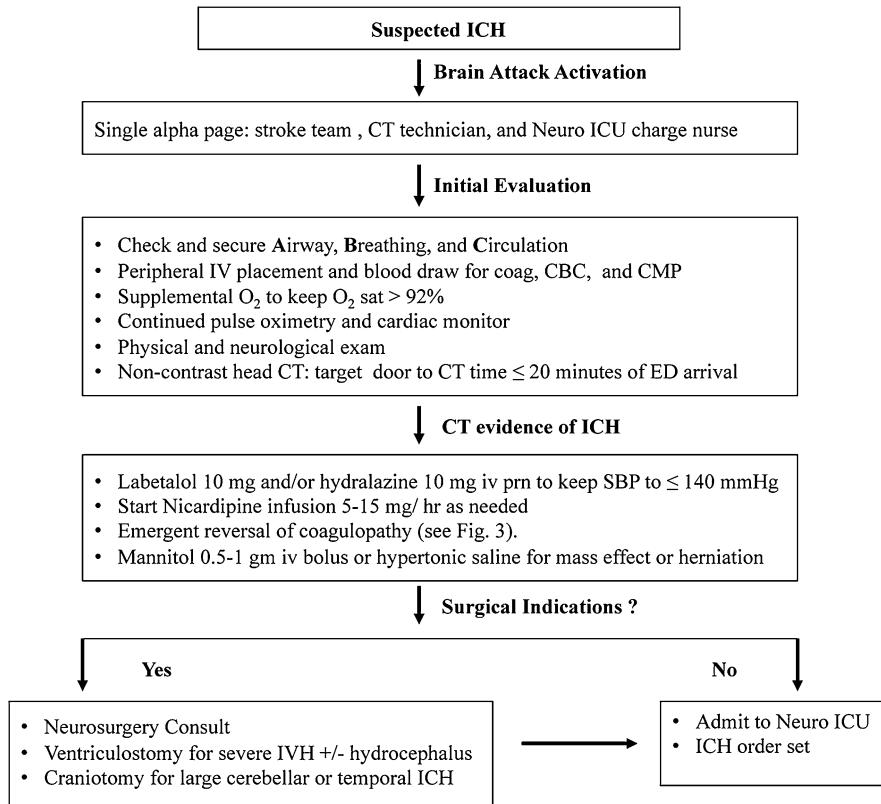


Fig. 5 Work flow for the management of intracerebral hemorrhage [112]

that the AI tools will soon become a valuable partner of the physicians. A physician's decision making would significantly improve if AI tools provide timely information that the physician would otherwise take a while to process himself. Information of the risk of drug interactions or side effects, patient suitability for reperfusion therapy, identification and prediction of stroke mechanism, and integration of wealth of clinical, proteomic, and genomic data to individualize patient care would be a great leap forward in the patient care. As an example, radiologists are confronted with a huge volume of imaging data to sift through. AIDOCs which runs in the background highlights the images (and regions therein) for the radiologist to give greater attention and thus prevent missing clinically important imaging feature.

An AI algorithm engineer needs a clearly defined ground truth based on a clinician's expertise; a clinician needs technical expertise to design algorithm suited to the clinical needs. A symbiotic relationship between the stroke researchers, AI

algorithm engineers, and regulatory bodies would be crucial to advance the use of AI in stroke medicine [114].

References

- Gorelick PB. The global burden of stroke: persistent and disabling. *Lancet Neurol*. 2019;18(5):417–8.
- Mikulik R, Wahlgren N. Treatment of acute stroke: an update. *J Intern Med*. 2015;278(2):145–65.
- Hemphill JC 3rd, Greenberg SM, Anderson CS, Becker K, Bendok BR, Cushman M, et al. Guidelines for the management of spontaneous intracerebral hemorrhage: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke*. 2015;46(7):2032–60.
- Schregel K, Behme D, Tsogkas I, Knauth M, Maier I, Karch A, et al. Effects of workflow optimization in endovascularly treated stroke patients – a pre-post effectiveness study. *PLoS One*. 2016;11(12):e0169192.
- Hacke W, Kaste M, Bluhmki E, Brozman M, Davalos A, Guidetti D, et al. Thrombolysis with alteplase 3 to 4.5 hours after acute ischemic stroke. *N Engl J Med*. 2008;359(13):1317–29.

6. Emberson J, Lees KR, Lyden P, Blackwell L, Albers G, Bluhmki E, et al. Effect of treatment delay, age, and stroke severity on the effects of intravenous thrombolysis with alteplase for acute ischaemic stroke: a meta-analysis of individual patient data from randomised trials. *Lancet*. 2014;384(9958):1929–35.
7. Hazinski M. D-mystifying recognition and management of stroke. *Curr Emerg Cardiac Care*. 1996;7:8.
8. Mishra NK, Patel H, Hastak SM. Comprehensive stroke care: an overview. *J Assoc Physicians India*. 2006;54:36–41.
9. Hacke W, Donnan G, Fieschi C, Kaste M, von Kummer R, Broderick JP, et al. Association of outcome with early stroke treatment: pooled analysis of ATLANTIS, ECASS, and NINDS rt-PA stroke trials. *Lancet*. 2004;363(9411):768–74.
10. National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. Tissue plasminogen activator for acute ischemic stroke. *N Engl J Med*. 1995;333(24):1581–7.
11. Berkhemer OA, Fransen PS, Beumer D, van den Berg LA, Lingsma HF, Yoo AJ, et al. A randomized trial of intraarterial treatment for acute ischemic stroke. *N Engl J Med*. 2015;372(1):11–20.
12. Goyal M, Demchuk AM, Menon BK, Eesa M, Rempel JL, Thornton J, et al. Randomized assessment of rapid endovascular treatment of ischemic stroke. *N Engl J Med*. 2015;372(11):1019–30.
13. Campbell BC, Mitchell PJ, Kleimig TJ, Dewey HM, Churilov L, Yassi N, et al. Endovascular therapy for ischemic stroke with perfusion-imaging selection. *N Engl J Med*. 2015;372(11):1009–18.
14. Goyal M, Menon BK, van Zwam WH, Dippel DW, Mitchell PJ, Demchuk AM, et al. Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials. *Lancet*. 2016;387(10029):1723–31.
15. Menon BK, Hill MD, Davalos A, Roos Y, Campbell BCV, Dippel DWJ, et al. Efficacy of endovascular thrombectomy in patients with M2 segment middle cerebral artery occlusions: meta-analysis of data from the HERMES Collaboration. *J Neurointerv Surg*. 2019;11(11):1065–9.
16. Jovin TG, Saver JL, Ribo M, Pereira V, Furlan A, Bonafe A, et al. Diffusion-weighted imaging or computerized tomography perfusion assessment with clinical mismatch in the triage of wake up and late presenting strokes undergoing neurointervention with Trevo (DAWN) trial methods. *Int J Stroke*. 2017;12(6):641–52.
17. Albers GW, Marks MP, Kemp S, Christensen S, Tsai JP, Ortega-Gutierrez S, et al. Thrombectomy for stroke at 6 to 16 hours with selection by perfusion imaging. *N Engl J Med*. 2018;378(8):708–18.
18. Aghaebrahim A, Granja MF, Agnoletto GJ, Aguilar-Salinas P, Cortez GM, Santos R, et al. Workflow optimization for ischemic stroke in a community-based stroke center. *World Neurosurg*. 2019;129:e273–e8.
19. Heo JH, Kim YD, Nam HS, Hong KS, Ahn SH, Cho HJ, et al. A computerized in-hospital alert system for thrombolysis in acute stroke. *Stroke*. 2010;41(9):1978–83.
20. Lee JH, Oh BJ, Ahn JY, Lee SW, Lee YH, Min SW, et al. Effectiveness of automatic acute stroke alert system based on UMLS mapped local terminology codes at emergency department. *AMIA Annu Symp Proc*. 2008;1018.
21. Seah HM, Burney M, Phan M, Shell D, Wu J, Zhou K, et al. CODE STROKE ALERT-concept and development of a novel open-source platform to streamline acute stroke management. *Front Neurol*. 2019;10:725.
22. Alabdali A, Yousif S, Alsaleem A, Aldhubayb M, Aljerian N. Can Emergency Medical Services (EMS) Shorten the Time to Stroke Team Activation, Computed Tomography (CT), and the Time to Receiving Antithrombotic Therapy? A prospective cohort study. *Prehosp Disaster Med*. 2020;35(2):148–51.
23. Fassbender K, Walter S, Grunwald IQ, Merzou F, Mathur S, Lesmeister M, et al. Prehospital stroke management in the thrombectomy era. *Lancet Neurol*. 2020;19(7):601–10.
24. Helwig SA, Ragoschke-Schumm A, Schwindling L, Kettner M, Roumia S, Kulikovski J, et al. Prehospital stroke management optimized by use of clinical scoring vs mobile stroke unit for triage of patients with stroke: a randomized clinical trial. *JAMA Neurol*. 2019;76:1484.
25. Li S, Wang A, Zhang X, Wang Y. Design and validation of prehospital acute stroke triage (PAST) scale to predict large vessel occlusion. *Atherosclerosis*. 2020;306:1–5.
26. Krebs S, Roth D, Knoflach M, Baubin M, Lang W, Beisteiner R, et al. Design and derivation of the Austrian Prehospital Stroke Scale (APSS) to predict severe stroke with large vessel occlusion. *Prehosp Emerg Care*. 2020;1:1–8.
27. Mazya MV, Berglund A, Ahmed N, von Euler M, Holmin S, Laska AC, et al. Implementation of a prehospital stroke triage system using symptom severity and teleconsultation in the stockholm stroke triage study. *JAMA Neurol*. 2020;77(6):691–9.
28. Baker DW, Tschurtz BA, Aliaga AE, Williams SC, Jauch EC, Schwamm LH. Determining the need for thrombectomy-capable stroke centers based on travel time to the nearest comprehensive stroke center. *Jt Comm J Qual Patient Saf*. 2020;46(9):501–5.
29. Holodinsky JK, Francis MJ, Goyal M, Hill MD, Kamal N. Testing the usability of a software for geospatial and transport modeling in acute stroke service planning. *Front Neurol*. 2019;10:694.
30. Tajaddini A, Phan TG, Beare R, Ma H, Srikanth V, Currie G, et al. Application of strategic transport model and Google maps to develop better clot retrieval stroke service. *Front Neurol*. 2019;10:692.
31. Newman-Toker DE, Moy E, Valente E, Coffey R, Hines AL. Missed diagnosis of stroke in the emergency department: a cross-sectional analysis of a large population-based sample. *Diagnosis (Berl)*. 2014;1(2):155–66.

32. Abedi V, Goyal N, Tsivgoulis G, HosseiniChimeh N, Hontecillas R, Bassaganya-Riera J, et al. Novel screening tool for stroke using artificial neural network. *Stroke.* 2017;48(6):1678–81.
33. Chen Z, Zhang R, Xu F, Gong X, Shi F, Zhang M, et al. Novel prehospital prediction model of large vessel occlusion using artificial neural network. *Front Aging Neurosci.* 2018;10:181.
34. Maas WJ, Lahr MMH, Buskens E, van der Zee DJ, Uyttenboogaart M, Investigators C. Pathway design for acute stroke care in the era of endovascular thrombectomy: a critical overview of optimization efforts. *Stroke.* 2020;51(11):3452–60.
35. Herpich F, Rincon F. Management of acute ischemic stroke. *Crit Care Med.* 2020;48(11):1654–63.
36. Powers WJ, Rabinstein AA, Ackerson T, Adeoye OM, Bambakidis NC, Becker K, et al. 2018 guidelines for the early management of patients with acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke.* 2018;49(3):e46–e110.
37. Turc G, Bhogal P, Fischer U, Khatri P, Lobotesis K, Mazighi M, et al. European Stroke Organisation (ESO)- European Society for Minimally Invasive Neurological Therapy (ESMINT) guidelines on mechanical thrombectomy in acute ischemic stroke. *J Neurointerv Surg.* 2019;11(6):535–8.
38. Frank B, Grotta JC, Alexandrov AV, Bluhmki E, Lyden P, Meretoja A, et al. Thrombolysis in stroke despite contraindications or warnings? *Stroke.* 2013;44(3):727–33.
39. Powers WJ, Rabinstein AA, Ackerson T, Adeoye OM, Bambakidis NC, Becker K, et al. Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke.* 2019;50(12):e344–418.
40. Thomalla G, Cheng B, Ebinger M, Hao Q, Tourdias T, Wu O, et al. DWI-FLAIR mismatch for the identification of patients with acute ischaemic stroke within 4.5 h of symptom onset (PRE-FLAIR): a multicentre observational study. *Lancet Neurol.* 2011;10(11):978–86.
41. Etherton MR, Barreto AD, Schwamm LH, Wu O. Neuroimaging paradigms to identify patients for reperfusion therapy in stroke of unknown onset. *Front Neurol.* 2018;9:327.
42. Ho KC, Speier W, El-Saden S, Arnold CW. Classifying acute ischemic stroke onset time using deep imaging features. *AMIA Annu Symp Proc.* 2017;2017:892–901.
43. Ho KC, Speier W, Zhang H, Scalzo F, El-Saden S, Arnold CW. A machine learning approach for classifying ischemic stroke onset time from imaging. *IEEE Trans Med Imaging.* 2019;38(7):1666–76.
44. Lees KR, Bluhmki E, von Kummer R, Brott TG, Toni D, Grotta JC, et al. Time to treatment with intravenous alteplase and outcome in stroke: an updated pooled analysis of ECASS, ATLANTIS, NINDS, and EPITHET trials. *Lancet.* 2010;375(9727):1695–703.
45. Marler JR, Tilley BC, Lu M, Brott TG, Lyden PC, Grotta JC, et al. Early stroke treatment associated with better outcome: the NINDS rt-PA stroke study. *Neurology.* 2000;55(11):1649–55.
46. Mishra NK, Albers GW, Davis SM, Donnan GA, Furlan AJ, Hacke W, et al. Mismatch-based delayed thrombolysis: a meta-analysis. *Stroke.* 2010;41(1):e25–33.
47. Mishra NK, Albers GW, Christensen S, Marks M, Hamilton S, Straka M, et al. Comparison of magnetic resonance imaging mismatch criteria to select patients for endovascular stroke therapy. *Stroke.* 2014;45(5):1369–74.
48. Trojaborg W, Boysen G. Relation between EEG, regional cerebral blood flow and internal carotid artery pressure during carotid endarterectomy. *Electroencephalogr Clin Neurophysiol.* 1973;34(1):61–9.
49. Sharbrough FW, Messick JM Jr, Sundt TM Jr. Correlation of continuous electroencephalograms with cerebral blood flow measurements during carotid endarterectomy. *Stroke.* 1973;4(4):674–83.
50. Sundt TM Jr, Sharbrough FW, Anderson RE, Michenfelder JD. Cerebral blood flow measurements and electroencephalograms during carotid endarterectomy. *J Neurosurg.* 1974;41(3):310–20.
51. Jones TH, Morawetz RB, Crowell RM, Marcoux FW, FitzGibbon SJ, DeGirolami U, et al. Thresholds of focal cerebral ischemia in awake monkeys. *J Neurosurg.* 1981;54(6):773–82.
52. Bell BA, Symon L, Branston NM. CBF and time thresholds for the formation of ischemic cerebral edema, and effect of reperfusion in baboons. *J Neurosurg.* 1985;62(1):31–41.
53. Morawetz RB, Jones TH, Ojemann RG, Marcoux FW, DeGirolami U, Crowell RM. Regional cerebral blood flow during temporary middle cerebral artery occlusion in waking monkeys. *Acta Neurol Scand Suppl.* 1977;64:114–5.
54. Marcoux FW, Morawetz RB, Crowell RM, DeGirolami U, Halsey JH Jr. Differential regional vulnerability in transient focal cerebral ischemia. *Stroke.* 1982;13(3):339–46.
55. Grotta JC, Alexandrov AV. tPA-associated reperfusion after acute stroke demonstrated by SPECT. *Stroke.* 1998;29(2):429–32.
56. Latchaw RE, Yonas H, Hunter GJ, Yuh WT, Ueda T, Sorensen AG, et al. Guidelines and recommendations for perfusion imaging in cerebral ischemia: a scientific statement for healthcare professionals by the writing group on perfusion imaging, from the Council on Cardiovascular Radiology of the American Heart Association. *Stroke.* 2003;34(4):1084–104.
57. Olivot JM, Mlynash M, Zaharchuk G, Straka M, Bammer R, Schwartz N, et al. Perfusion MRI (Tmax

- and MTT) correlation with xenon CT cerebral blood flow in stroke patients. *Neurology*. 2009;72(13):1140–5.
58. Takasawa M, Jones PS, Guadagno JV, Christensen S, Fryer TD, Harding S, et al. How reliable is perfusion MR in acute stroke? Validation and determination of the penumbra threshold against quantitative PET. *Stroke*. 2008;39(3):870–7.
 59. Calamante F, Connelly A, van Osch MJ. Nonlinear DeltaR*2 effects in perfusion quantification using bolus-tracking MRI. *Magn Reson Med*. 2009;61(2):486–92.
 60. Calamante F, Gadian DG, Connelly A. Quantification of perfusion using bolus tracking magnetic resonance imaging in stroke: assumptions, limitations, and potential implications for clinical use. *Stroke*. 2002;33(4):1146–51.
 61. Calamante F, Yim PJ, Cebral JR. Estimation of bolus dispersion effects in perfusion MRI using image-based computational fluid dynamics. *NeuroImage*. 2003;19(2 Pt 1):341–53.
 62. Willats L, Christensen S, Ma HK, Donnan GA, Connelly A, Calamante F. Validating a local Arterial Input Function method for improved perfusion quantification in stroke. *J Cereb Blood Flow Metab*. 2011;31(11):2189–98.
 63. Straka M, Albers GW, Bammer R. Real-time diffusion-perfusion mismatch analysis in acute stroke. *J Magn Reson Imaging*. 2010;32(5):1024–37.
 64. Chen C, Bivard A, Lin L, Levi CR, Spratt NJ, Parsons MW. Thresholds for infarction vary between gray matter and white matter in acute ischemic stroke: a CT perfusion study. *J Cereb Blood Flow Metab*. 2019;39(3):536–46.
 65. Calamante F. Arterial input function in perfusion MRI: a comprehensive review. *Prog Nucl Magn Reson Spectrosc*. 2013;74:1–32.
 66. Calamante F. Bolus dispersion issues related to the quantification of perfusion MRI data. *J Magn Reson Imaging*. 2005;22(6):718–22.
 67. Calamante F, Willats L, Gadian DG, Connelly A. Bolus delay and dispersion in perfusion MRI: implications for tissue predictor models in stroke. *Magn Reson Med*. 2006;55(5):1180–5.
 68. Calamante F, Morup M, Hansen LK. Defining a local arterial input function for perfusion MRI using independent component analysis. *Magn Reson Med*. 2004;52(4):789–97.
 69. Ho KC, Scalzo F, Sarma KV, Speier W, El-Saden S, Arnold C. Predicting ischemic stroke tissue fate using a deep convolutional neural network on source magnetic resonance perfusion images. *J Med Imag (Bellingham)*. 2019;6(2):026001.
 70. Olivot JM, Mlynash M, Thijs VN, Kemp S, Lansberg MG, Wechsler L, et al. Optimal Tmax threshold for predicting penumbral tissue in acute stroke. *Stroke*. 2009;40(2):469–75.
 71. Kakuda W, Lansberg MG, Thijs VN, Kemp SM, Bammer R, Wechsler LR, et al. Optimal definition for PWI/DWI mismatch in acute ischemic stroke patients. *J Cereb Blood Flow Metab*. 2008;28(5):887–91.
 72. Albers GW, Thijs VN, Wechsler L, Kemp S, Schlaug G, Skalabrin E, et al. Magnetic resonance imaging profiles predict clinical response to early reperfusion: the diffusion and perfusion imaging evaluation for understanding stroke evolution (DEFUSE) study. *Ann Neurol*. 2006;60(5):508–17.
 73. Lansberg MG, Straka M, Kemp S, Mlynash M, Wechsler LR, Jovic TG, et al. MRI profile and response to endovascular reperfusion after stroke (DEFUSE 2): a prospective cohort study. *Lancet Neurol*. 2012;11(10):860–7.
 74. Lansberg MG, Christensen S, Kemp S, Mlynash M, Mishra N, Federau C, et al. Computed tomographic perfusion to predict response to recanalization in ischemic stroke. *Ann Neurol*. 2017;81(6):849–56.
 75. Cereda CW, Christensen S, Campbell BCV, Mishra NK, Mlynash M, Levi C, et al. A benchmarking tool to evaluate computer tomography perfusion infarct core predictions against a DWI standard. *J Cereb Blood Flow Metab*. 2016;36(10):1780–9.
 76. Soize S, Tisserand M, Charron S, Turc G, Ben Hassen W, Labeyrie MA, et al. How sustained is 24-hour diffusion-weighted imaging lesion reversal? Serial magnetic resonance imaging in a patient cohort thrombolyzed within 4.5 hours of stroke onset. *Stroke*. 2015;46(3):704–10.
 77. Inoue M, Mlynash M, Christensen S, Wheeler HM, Straka M, Tipirneni A, et al. Early diffusion-weighted imaging reversal after endovascular reperfusion is typically transient in patients imaged 3 to 6 hours after onset. *Stroke*. 2014;45(4):1024–8.
 78. Campbell BC, Purushotham A, Christensen S, Desmond PM, Nagakane Y, Parsons MW, et al. The infarct core is well represented by the acute diffusion lesion: sustained reversal is infrequent. *J Cereb Blood Flow Metab*. 2012;32(1):50–6.
 79. Chemmanam T, Campbell BC, Christensen S, Nagakane Y, Desmond PM, Bladin CF, et al. Ischemic diffusion lesion reversal is uncommon and rarely alters perfusion-diffusion mismatch. *Neurology*. 2010;75(12):1040–7.
 80. Olivot JM, Mlynash M, Thijs VN, Purushotham A, Kemp S, Lansberg MG, et al. Relationships between cerebral perfusion and reversibility of acute diffusion lesions in DEFUSE: insights from RADAR. *Stroke*. 2009;40(5):1692–7.
 81. Chen L, Bentley P, Rueckert D. Fully automatic acute ischemic lesion segmentation in DWI using convolutional neural networks. *Neuroimage Clin*. 2017;15:633–43.
 82. Mishra NK, Christensen S, Wouters A, Campbell BC, Straka M, Mlynash M, et al. Reperfusion of very low cerebral blood volume lesion predicts parenchymal hematoma after endovascular therapy. *Stroke*. 2015;46(5):1245–9.

83. Bang OY, Buck BH, Saver JL, Alger JR, Yoon SR, Starkman S, et al. Prediction of hemorrhagic transformation after recanalization therapy using T2*-permeability magnetic resonance imaging. *Ann Neurol.* 2007;62(2):170–6.
84. Bang OY, Saver JL, Alger JR, Shah SH, Buck BH, Starkman S, et al. Patterns and predictors of blood-brain barrier permeability derangements in acute ischemic stroke. *Stroke.* 2009;40(2):454–61.
85. Yassi N, Parsons MW, Christensen S, Sharma G, Bivard A, Donnan GA, et al. Prediction of poststroke hemorrhagic transformation using computed tomography perfusion. *Stroke.* 2013;44(11):3039–43.
86. Lemmens R, Hamilton SA, Liebeskind DS, Tomsick TA, Demchuk AM, Nogueira RG, et al. Effect of endovascular reperfusion in relation to site of arterial occlusion. *Neurology.* 2016;86(8):762–70.
87. Amukotuwa SA, Straka M, Smith H, Chandra RV, Dehkharhani S, Fischbein NJ, et al. Automated detection of intracranial large vessel occlusions on computed tomography angiography: a single center experience. *Stroke.* 2019;50(10):2790–8.
88. Amukotuwa SA, Straka M, Dehkharhani S, Bammer R. Fast automatic detection of large vessel occlusions on CT angiography. *Stroke.* 2019;50(12):3431–8.
89. Benson JC, Fitzgerald ST, Kadivel R, Johnson C, Dai D, Karen D, et al. Clot permeability and histopathology: is a clot's perviousness on CT imaging correlated with its histologic composition? *J Neurointerv Surg.* 2020;12(1):38–42.
90. Berndt M, Muck F, Maegerlein C, Wunderlich S, Zimmer C, Wirth S, et al. Introduction of CTA-index as simplified measuring method for Thrombus perviousness. *Clin Neuroradiol.* 2020.
91. Santos EM, Marquering HA, den Blanken MD, Berkhemer OA, Boers AM, Yoo AJ, et al. Thrombus permeability is associated with improved functional outcome and recanalization in patients with ischemic stroke. *Stroke.* 2016;47(3):732–41.
92. Borst J, Berkhemer OA, Santos EMM, Yoo AJ, den Blanken M, Roos Y, et al. Value of thrombus CT characteristics in patients with acute ischemic stroke. *AJNR Am J Neuroradiol.* 2017;38(9):1758–64.
93. Alves HC, Treurniet KM, Dutra BG, Jansen IGH, Boers AMM, Santos EMM, et al. Associations between collateral status and thrombus characteristics and their impact in anterior circulation stroke. *Stroke.* 2018;49(2):391–6.
94. Bilgic AB, Gocmen R, Arsava EM, Topcuoglu MA. The effect of clot volume and permeability on response to intravenous tissue plasminogen activator in acute ischemic stroke. *J Stroke Cerebrovasc Dis.* 2020;29(2):104541.
95. Adams HP Jr, Bendixen BH, Kappelle LJ, Biller J, Love BB, Gordon DL, et al. Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. *Stroke.* 1993;24(1):35–41.
96. Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet.* 2019;394(10201):861–7.
97. Kashou AH, Rabinstein AA, Attia IZ, Asirvatham SJ, Gersh BJ, Friedman PA, et al. Recurrent cryptogenic stroke: a potential role for an artificial intelligence-enabled electrocardiogram? *HeartRhythm Case Rep.* 2020;6(4):202–5.
98. Albers GW, Amarenco P, Easton JD, Sacco RL, Teal P. Antithrombotic and thrombolytic therapy for ischemic stroke: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines (8th Edition). *Chest.* 2008;133(6 Suppl):630S–69S.
99. Diener HC, Bernstein R, Hart R. Secondary stroke prevention in cryptogenic stroke and embolic stroke of undetermined source (ESUS). *Curr Neurol Neurosci Rep.* 2017;17(9):64.
100. Schabitz WR, Kohrmann M, Schellinger PD, Minnerup J, Fisher M. Embolic stroke of undetermined source: gateway to a new stroke entity? *Am J Med.* 2020;133(7):795–801.
101. Hart RG, Catanese L, Perera KS, Ntaios G, Connolly SJ. Embolic stroke of undetermined source: a systematic review and clinical update. *Stroke.* 2017;48(4):867–72.
102. Ntaios G, Weng SF, Perlepe K, Akyea R, Condon L, Lambrou D, et al. Data-driven machine-learning analysis of potential embolic sources in embolic stroke of undetermined source. *Eur J Neurol.* 2020.
103. Kamel H, Navi BB, Parikh NS, Merkler AE, Okin PM, Devereux RB, et al. Machine learning prediction of stroke mechanism in embolic strokes of undetermined source. *Stroke.* 2020;51(9):e203–e10.
104. Rothwell PM, Giles MF, Flossmann E, Lovelock CE, Redgrave JN, Warlow CP, et al. A simple score (ABCD) to identify individuals at high early risk of stroke after transient ischaemic attack. *Lancet.* 2005;366(9479):29–36.
105. Giles MF, Albers GW, Amarenco P, Arsava MM, Asimos A, Ay H, et al. Addition of brain infarction to the ABCD2 Score (ABCD2I): a collaborative analysis of unpublished data on 4574 patients. *Stroke.* 2010;41(9):1907–13.
106. Zaharchuk G, Olivot JM, Fischbein NJ, Bammer R, Straka M, Kleinman JT, et al. Arterial spin labeling imaging findings in transient ischemic attack patients: comparison with diffusion- and bolus perfusion-weighted imaging. *Cerebrovasc Dis.* 2012;34(3):221–8.
107. Easton JD, Saver JL, Albers GW, Alberts MJ, Chaturvedi S, Feldmann E, et al. Definition and evaluation of transient ischemic attack: a scientific statement for healthcare professionals from the American Heart Association/American Stroke Association Stroke Council; Council on Cardiovascular Surgery and Anesthesia; Council on Cardiovascular Radiology and Intervention; Council on Cardiovascular Nursing; and Council on Stroke. *Stroke.* 2014;45(3):879–908.

- and the Interdisciplinary Council on Peripheral Vascular Disease. The American Academy of Neurology affirms the value of this statement as an educational tool for neurologists. *Stroke.* 2009;40(6):2276–93.
108. Kleinman JT, Zaharchuk G, Mlynash M, Ogdie AA, Straka M, Lansberg MG, et al. Automated perfusion imaging for the evaluation of transient ischemic attack. *Stroke.* 2012;43(6):1556–60.
109. Mlynash M, Olivot JM, Tong DC, Lansberg MG, Eynsorn I, Kemp S, et al. Yield of combined perfusion and diffusion MR imaging in hemispheric TIA. *Neurology.* 2009;72(13):1127–33.
110. Bacchi S, Oakden-Rayner L, Zerner T, Kleinig T, Patel S, Jannes J. Deep learning natural language processing successfully predicts the cerebrovascular cause of transient ischemic attack-like presentations. *Stroke.* 2019;50(3):758–60.
111. Steiner T, Al-Shahi Salman R, Beer R, Christensen H, Cordonnier C, Csiba L, et al. European Stroke Organisation (ESO) guidelines for the management of spontaneous intracerebral hemorrhage. *Int J Stroke.* 2014;9(7):840–55.
112. Dastur CK, Yu W. Current management of spontaneous intracerebral haemorrhage. *Stroke Vasc Neurol.* 2017;2(1):21–9.
113. Liu J, Xu H, Chen Q, Zhang T, Sheng W, Huang Q, et al. Prediction of hematoma expansion in spontaneous intracerebral hemorrhage using support vector machine. *EBioMedicine.* 2019;43: 454–9.
114. Food and Drug Administration, Policy for Device Software Functions and Mobile Medical Applications, 2019. Weblink: <https://www.fda.gov/media/80958/download>



AIM in Clinical Neurophysiology and Electroencephalography (EEG)

125

Joseph Davids, Viraj Bharambe, and Hutan Ashrafian

Contents

Introduction	1754
Epilepsy	1755
Artificial Intelligence, Machine Learning, and Deep Learning	1755
Machine Learning and Deep Learning Approaches in Epilepsy	1757
The Role of History-Taking and Where Deep Learning Can Make In-Roads	1757
Deep Learning Approaches for Investigating Epilepsy	1758
Machine Learning for EEG Analysis	1758
Deep Learning in Epilepsy Treatment	1759
Machine Learning for Epilepsy Surgery	1761

Joseph Davids and Viraj Bharambe are joint first authors

J. Davids (✉)
Imperial College London NHS Trust, London, UK

Hamlyn Centre for Robotics and Artificial Intelligence,
Department of Surgery and Cancer, Imperial College
London, London, UK

National Hospital for Neurology and Neurosurgery Queen
Square, London, UK
e-mail: j.davids@imperial.ac.uk

V. Bharambe
Walton Centre for Neurology and Neurosurgery,
Liverpool, UK

H. Ashrafian
Department of Surgery and Cancer, Imperial College
London NHS Trust, London, UK

Institute of Global Health Innovation, Imperial College
London, London, UK

Hamlyn Centre for Robotics and Artificial Intelligence,
Department of Surgery and Cancer, Imperial College
London, London, UK

Machine Learning for Electrophysiological Migraine Detection	1762
Deep Learning for Electromyography	1762
Practical Machine Learning for EEG Analysis	1763
Concluding Remarks	1764
Highlighting the Tension Between Progress and “Model Explainability”	1764
References	1764

Abstract

Artificial intelligence and its facets of machine and deep learning have permeated into the fabric of our society with widespread adoption. However, the medical community is only recently beginning to embrace its potential for applications in various subspecialties. Clinical neurophysiology, neurology, neurosurgery, and the rest of the neurosciences are seeing considerable impacts in how AI is being utilized across these specialties. EEG analysis using deep learning in the field of neuro-pathophysiology is also gaining rapid traction with a myriad of emerging applications suggesting promise. However, although deep learning has shown considerable potential, the lack of transparency about how models make decisions and the barriers to entry curtail acceptability among clinicians. Some have argued that the stakes for erroneous judgments in the diagnostic pathway remain too high to be completely reliant on AI, while others have embraced its potential and are delivering services using AI. This chapter discusses the role of AI in clinical neurophysiology with an extended focus on epilepsy and EEG analysis. We also highlight some of the areas where AI and its applications have been adopted for characterization of other neurophysiological diagnostic modalities, for instance, in migraine, and end with a discussion of model explainability.

Keywords

Epilepsy · EEG · Clinical neurophysiology · Deep learning · Seizures · AI · Machine learning · Artificial intelligence

Introduction

As a subspecialty of clinical neurosciences, neurophysiology provides theranostic avenues for patients with a myriad of neurological diseases ranging from peripheral nervous system disorders, disorders of muscle, and nerve conduction abnormalities to central nervous system disorders such as Parkinson’s disease, spinal cord pathology, and epilepsy.

Traditionally, investigations in clinical neurophysiology have fallen into two facets: (a) the study of brain function using electroencephalography, on the scalp or direct to brain contact electrode recordings of the intrinsic electrical neuronal activity, and (b) the study of peripheral nervous system disorders using electromyography and nerve conduction studies. The latter can provide precise information about neurodegeneration or neuronal dysfunction at the ion channel, neuromuscular junction, or corticospinal tract level. While the former can precisely identify aspects of intracranial neuronal dysfunction at millisecond resolution and correlate it to brain structure, these investigations are still supplementary adjuncts to the patient’s history and examination, which form the basis of all neurological diagnosis.

Over the past 100 years, this subspecialty has also evolved into supporting intraoperative monitoring of neuroanatomical pathways during epilepsy surgery, Parkinson’s disease surgery, spinal cord surgery, and brain tumor excision. Functional monitoring and brain mapping are used to alert or forewarn the surgeon of likely iatrogenic injury to eloquent areas within the central nervous system. Monitoring establishes baseline amplitude, latencies and frequencies, or the presence of discharges to

enable ground truth comparison with abnormal changes in these areas, whereas testing localizes normal functioning eloquent tissue and also differentiates it from dysfunctional tissue for surgical resection [1].

Similarly, over the past 50 years, since Turing and collaborators hinted at intelligent machines within the field of AI, Hinton and others have made important contributions with deep learning [2]. It has seen widespread adoption in various medical specialties, as one can read in the respective chapters. Over the past decade, the field has seen unprecedented growth in applications of AI for neuroscience specialties as well.

The focus of this chapter will primarily be on functional brain monitoring and diagnostics using electroencephalography and the applications of AI and machine learning in this area, with an additional focus on epilepsy. However, we will also briefly identify aspects of brain mapping where progress has also been made using machine learning.

Epilepsy

Epilepsy is defined as a disorder of the brain characterized by an enduring predisposition to generate epileptiform seizures and the neurobiological, cognitive, psychological and social consequences of this condition [3]. Epilepsy is the term used to describe a recurrent seizure disorder.

An epileptic seizure is the transient occurrence of signs and/or symptoms due to abnormal excessive or synchronous neuronal activity in the brain [3]. Seizures and epilepsy are common: 1 in 20 people will experience a single seizure at some point in their life and 1 in 50 will have epilepsy [4].

The approach to any patient presenting with epilepsy is to classify it as accurately as possible as this has a significant impact on the approach to treatment. The current approach to diagnosis operates at three levels: seizure type(s), epilepsy type(s), and epilepsy syndrome. Where possible, a diagnosis at all three levels should be

sought as well as the etiology of the individual's epilepsy [5].

The first diagnostic quandary for any patient presenting with seizure(s) is to classify whether the problem is focal or generalized.

The term "focal-onset epilepsy" is best understood in its unifocal variation. Unifocal-onset epilepsy indicates that a person's seizures arise from a focal lesion within the brain matter and then may or may not spread to other brain regions. The epileptogenic lesion can be caused by a wide variety of insults including infection, trauma, tumors, and strokes to name a few. The lesion produces paroxysmal, stereotyped signs and symptoms which can then progress on to a generalized seizure if the abnormal neuronal activity spreads.

The diagnosis of "generalized epilepsy" is made on clinical grounds and a clinician would have to identify that the individual experiences a range of seizure types including absence, myoclonic, atonic, tonic, and tonic-clonic. Importantly, the patient's electroencephalogram (EEG) would typically show generalized spike-wave activity and support the diagnosis [5].

This chapter will focus on the diagnostic challenges that clinicians and patients face at various stages along the patient pathway and the ways in which AI/machine learning could support this.

Artificial Intelligence, Machine Learning, and Deep Learning

Artificial intelligence is an umbrella field encompassing machine learning and deep learning. Machine learning combines statistics and computer science to develop algorithms. These algorithms are used on large meaningful datasets and refine themselves to improve performance. Machine learning paradigms can be divided into supervised, unsupervised, semi-supervised, transfer, and reinforcement learning paradigms. However, the algorithms familiar to most machine learning scientists usually leverage supervised, unsupervised, and reinforcement approaches for their training.

Various aspects of machine/deep learning are described below, but readers are encouraged to also refer to the earlier chapters of the book where we summarize some of these concepts. In the context of neurophysiology, supervised learning involves training an algorithm that maps a function containing an input of neurophysiological datasets to an output space by using, for example, labeled ground truth training data that is provided directly to the algorithm. This labeling is usually performed by an expert clinician or scientist forming the ground truth. Unsupervised learning algorithms, on the other hand, draw inferences from input neurophysiological datasets without any need for labeling. The data can be structured or unstructured and the most common approaches have been clustering and dimensionality reduction methods. We summarize both approaches below. However, general data science best practice principles should also be implemented from the data mining perspective, including clinical neuro-electrophysiological data preprocessing and cleaning, data integration and data transformation, etc.

The steps for a supervised approach vary but usually require the following:

- Acquisition of adequate well-prepared training data based on the preference of the researcher or clinician and based on the question they are trying to answer.
- The training set is sampled using the appropriate sampling methodology that must be representative of the current data universe of interest.
- The input features that are to be represented could either be feature-engineered into an appropriate vector space or in some instances could require dimensionality reduction to enable optimized training and improved model accuracy. Poly-dimensional datasets are common in neuro-electrophysiology.
 - Note that this step of dimensionality reduction is an unsupervised approach, but it is sometimes required for optimal supervised learning in a multidimensional dataset.
- The feature set, which may sometimes also require scaling or normalization as some of

the algorithms may need this for optimal learning.

- Selection from an appropriate class of models and in some cases a universe/an ensemble of models or designing the right learning model architecture that would be optimal for framing the dataset.
- Hyperparameter optimization or tuning of the model where specific hyperparameters are tweaked to optimize the output to a desired level of accuracy and to identify the best bias-variance trade-offs.
- Model accuracy evaluation and metric reporting.
- Production deployment of the model for inference.
- Designing and building a user-friendly dashboard or interface for user interactivity to apply a trained model to solve new problems.

Some of the steps above are in some cases combined into frameworks that abstract away the complexities to the basic steps of data preprocessing, train-test splitting, model selection, model training, and inference reporting.

The unsupervised approach usually requires no labeled data and uses self-organization to identify latent features within the dataset that may not be easily discernible or apparent to the researcher.

The approach nonexclusively includes the following:

- Data acquisition using an appropriate sampling methodology as explained above.
- Principal component analysis, K-best, or other dimensionality reduction analysis on the dataset to limit the “curse of multidimensionality” as mentioned in machine learning parlance.
- Eigenvector decomposition analysis on the covariance matrix of features given to the model.
- Feature vectorization and scaling methods used are also similar to the above supervised model-dependent learning approach.
- Model learning of a new feature space that captures the characterization of the original unlabeled feature matrix.

Table 1 Summarizes some examples of supervised and unsupervised learning algorithms. Some have both regression and classification derivatives

Supervised	Unsupervised
Common regression algorithms Linear Logistic regression Ridge regression	Clustering K-means clustering Hierarchical clustering OPTICS DBScan
Support vector machines Support vector regression Support vector classifier	Anomaly detection Isolation forest Local outlier factor
Decision trees and random forests Decision tree classifier Decision tree regressor Random forest regressor Random forest classifier	Neural networks Autoencoders Restricted Boltzmann machine Deep belief networks Self-organizing maps Hebbian learning Generative adversarial networks
Ensemble and boosting models Gradient boosting regression and classification Extra trees regression and classification XGBoost, CatBoost, and AdaBoost regression and classification	Latent variable modeling and dimensionality reduction Expectation maximization Principal components analysis

- Model accuracy evaluation and metric reporting.
- Production level deployment and inference.

Table 1 provides a brief summary of some machine learning model examples that have been used for both supervised and unsupervised learning paradigms.

Machine Learning and Deep Learning Approaches in Epilepsy

The Role of History-Taking and Where Deep Learning Can Make In-Roads

The foundation of epilepsy and other neurological diagnoses lies in a detailed account of the patient's medical history from the patient or their carer. A thorough profile is required consisting of the patient's demographics, age of onset, family history, seizure semiology, seizure frequency, triggers, risk factors, drug history, driving status, and substance misuse. Before the use of imaging or EEG, a clinician uses these inputs to judge whether a patient's given events and history are suggestive of a seizure/epilepsy. A pretest

diagnosis is generated, which the investigations then support or refute. Often, patient history carries the greatest weight in coming to a final diagnosis. This is because, in many circumstances, imaging and EEG return normal results. In focal epilepsy caused by small lesions, imaging with modern 3-tesla magnetic resonance scanners does not have enough resolution to identify them and EEG is only helpful if a recording is made during a seizure [6]. Thus, many patients with a classical history of seizures are often formally diagnosed and started on medical treatment even though their investigations are normal [7].

This foundational aspect of patient assessment is the one part of the patient journey where deep learning processes have seldom been applied. This is not surprising. In comparison to imaging and EEG, this person-to-person interaction is inherently more difficult to encode into the required datasets necessary for deep learning to be effective. Nevertheless, there may be a role for screening patients using pre-clinic questionnaires.

For example, take a patient who presents to primary healthcare services with transient loss of consciousness. The current method by which patients are referred to First Fit Clinics is time-consuming for primary healthcare practitioners

and results in a heterogenous group of etiologies. Recent studies suggest that 17–35% of patients referred to a First Seizure Clinic do not have epilepsy and could be better served in other specialty clinics [8, 9]. Instead of a non-specialist taking a curtailed history in a time-pressured environment, the general practitioner could point the patient/carer in the direction of an online questionnaire. The responses to the questionnaire would then be fed into a predictive model, which has been trained on a previously encoded dataset. The model could judge from these responses whether the patient's events fall into one of three categories (vasovagal syncope, cardiogenic syncope, or seizure) and advise on the best referral pathway. This could greatly decrease the burden on primary and acute healthcare services and decrease the time taken for patients to see the correct specialist for their complaint.

Deep Learning Approaches for Investigating Epilepsy

As deep learning analyzes large datasets, its utility in interpreting investigative material for epilepsy patients has been of growing interest. The areas of interest include seizure detection, neuroimaging analysis, and EEG analysis.

Seizure detection/prediction is of great interest as it promises to solve multiple problems for clinicians and patients. As mentioned above, a major challenge in diagnosis is the fact that seizures can be infrequent and vast resources are required to constantly record and interpret video-EEG data. Time-locked video-EEG is the gold standard for diagnosing seizures and requires multiple days of scalp EEG and at least one video feed. Currently in the UK, this means admitting a patient to a video telemetry room for 3–5 days and all the costs that this incurs. Automated seizure detection could reduce the need for an expert reviewer and lower diagnostic and treatment costs.

Novel approaches have been used in both adult and pediatric settings. Karayiannis et al. examined limb movements in bedside video recordings, training neural networks to classify recordings as focal clonic seizures, myoclonic seizures, or non-seizure movements [10]. They report seizure

detection sensitivities of 85.5–94.4% and specificities of 92.5–97.9% on a matched testing set. Other groups have added EEG data to video feed with incremental gains on diagnostic accuracy (sensitivity 96.2%, specificity 94.2%) [11]. The former studies are impressive not just because of the accuracy of the algorithm but due to the fact that this could be achieved on video data alone.

While video recording is cheap, accessible, and reliable when recording infants, adults can be a more difficult cohort as they often live lives outside of the camera's view. Accelerometers are cheap and robust and, when paired with a simple surface electromyography electrode, can reach sensitivity of 90.9% and clinical latency of 10.5 s [12]. This promises clear benefits for carers as they could be alerted each time the system detects a possible seizure.

Image analysis constitutes an important aspect of diagnosis and preparation for surgery in focal epilepsies. Machine learning has been deployed in the interpretation of various imaging modalities including magnetic resonance imaging (MRI), fluorodeoxyglucose positron emission tomography (FDG-PET), and diffusion tensor imaging (DTI). Of greatest interest to most healthcare systems are the MRI studies, as they are the most readily available. Of course, one would not require the assistance of advanced neural networks in locating large obvious lesions. The main utility would be in identifying small epileptogenic lesions often missed by human experts. El Azami et al. trained a neural network to identify heterotopias and blurring of the gray-white matter junction as outliers on T1-weighted images from 77 healthy controls. The testing run on 11 patient scans attained a sensitivity of 77% with a mean of 3.2 false-positive detections per patient, compared to a sensitivity of 54% and an average of 6.3 false positives using state-of-the-art gold standard statistical parametric mapping (Wellcome Centre for Human Neuroimaging, London, UK) [13].

Machine Learning for EEG Analysis

EEG records electrical activity of the brain and usually requires specific electrode placement according to international 10–20 electrode

Machine Learning for EEG analysis

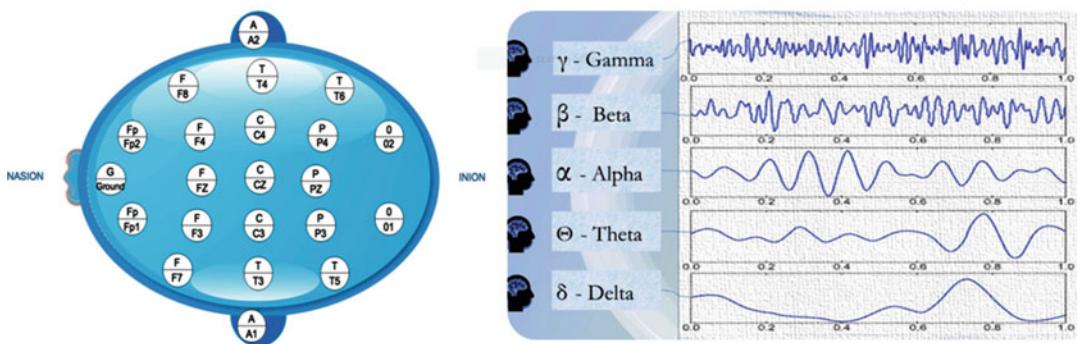


Fig. 1 Diagram of a single plane projection with electroencephalographic electrode placement. Left is the topographical map of a 21 EEG channel subset of international 10–20 placement common in all recordings and included in TUH Abnormal EEG Corpus (v2.0.0) first presented in the 1958 by Jasper [14, 15]. Right is the main characteristic EEG waveforms classified based on

placement criteria (see Fig. 1) [15]. Optimal electrode placement involves identifying landmarks such as the nasion, the inion, and the preauricular points that will provide the best signal-to-noise ratio.

- The electrodes can be placed in various, but specific, regions on the head and must have good contact with the skin of the scalp to limit impedance. Placement usually requires experience.
- Occasionally, sphenoidal electrodes are placed to enable deep structural recordings.
- There is also a routine need to reference the heart's rhythm through an ECG.

Based on placement, the EEG can be monitored using a montage-based system to generate functional outputs.

Montages include the following:

1. Bipolar montage
2. Referential montage
3. Average reference montage
4. Laplacian montage

An in-depth review of these approaches is beyond the scope of this chapter, but readers are referred to Siuly et al. [16]. For a physiologically normal adult, the reported amplitude of an

frequency and morphology. Gamma waveforms predominate during brain concentration and problem-solving. Beta activity reflects high brain activity or anxious states. Alpha arises during restful reflective states. Theta predominates during drowsiness. Delta is the asleep state. F, frontal; Fp, frontopolar; T, temporal; C, central; P, parietal; O, occipital; A, auricular (ear electrode)

EEG signal ranges from 1 to 100 μ V, and it is approximately 10–20 mV. The various detectable waveforms that have been characterized are:

- Delta with a frequency of ≤ 3 Hz
- Theta with a frequency between 3.5 and 7.5 Hz
- Alpha with a frequency between 7.5 and 13 Hz
- Beta with a frequency of ≥ 14 Hz
- Also, gamma oscillations >30 Hz

Machine learning could be used to study the optimal electrode placement when designing the EEG electrode cap and to educate trainees to be familiar with placement. This, together with sampling frequencies and other aspects of EEG recordings, could benefit from optimized and patient-targeted machine learning approaches (Table 2).

Deep Learning in Epilepsy Treatment

Across much of the world, most newly diagnosed patients with epilepsy are treated by primary healthcare practitioners. Some of these patients will respond to the first anti-epileptic drug (AED) they are prescribed, but those that do not are referred to a general neurologist. After the trial of the first drug, we know that further medication

Table 2 Adapted from Gemein et al. summarizing recent approaches in artificial intelligence on pathological EEG data discovery and the accuracies reached from applying various machine learning algorithms. CNN stands for

convolutional neural network and MLP stands for multi-layer perceptron (this is the technical term for artificial neural network); SVM stands for support vector machine [15]

Author	AI architecture	Data used	Accuracy reported [%]	Reference
Lopez de Diego et al. (2017)	CNN + MLP	1387 normal and 1398 abnormal files and an evaluation set of 150 normal and 130 abnormal files	78.8	[17]
Schirrmeyer et al. (2017a)	BD-Deep4	Filter banks common spatial pattern dataset, which underwent bandpass filtering, epoching, CSP computation, spatial filtering, feature construction and classification	85.4	[18]
Roy et al. (2019a)	ChronoNet	Temple University Hospital (TUH) Abnormal EEG Corpus dataset	86.6	[19, 20]
Amin et al. (2019)	AlexNet + SVM	Temple University Hospital (TUH) EEG Abnormal Corpus v2.0.0. Two classes: normal and abnormal. Evaluation set, the normal class = 148 subjects, and the abnormal class = 105 subjects. Training set, the normal class = 1237 subjects; the abnormal class = 893 subjects.	87.3	[20, 21]
Alhussein et al. (2019)	3 × AlexNet + MLP	EEG signals are made up of many frequency components. Researchers divided the frequency regions into several frequency bands called delta (1–3 Hz), theta (4–7 Hz), alpha (8–12 Hz), beta (12–30 Hz), and gamma (30–100 Hz)	89.1	[22]
Van Leeuwen et al. (2019)	BD-Deep4	Database of 8522 routine EEGs from the Department of Neurology in Massachusetts General Hospital collected over a 4-year period from 2012 to 2016	82.0	[23]

and polypharmacy are unlikely to achieve seizure freedom. It usually takes a trial of 3–4 AEDs before a patient is considered for epilepsy surgery. This sequential process means critical time is lost before drug-resistant patients can be assessed by epilepsy specialists for advanced treatment such as surgery. The associated time delay means such treatments may be less effective [24]. This can often result in years of reduced quality of life, lost productivity, and increased mortality [25].

An et al. compared machine learning algorithms for prediction of drug-resistant epilepsy (defined as requiring more than three medication changes during the study period) utilizing comprehensive US claims data from 2006 to 2015. They found that the best performing algorithm, trained using 635 features (comprising demographic variables, comorbidities, treatment regimens, insurance data, and clinical encounters) from 175,735 records, yielded an AUC of 76.4%

and could identify patients with drug-resistant epilepsy an average of 1.97 years before failing a second medication trial, using data available at the time of the first medication prescription [26].

After the first AED, patients face a trial-and-error-based approach to further medical management. This is often frustrating and time-consuming and carries a risk of hospitalization and mortality. Many studies have demonstrated the possibility of using machine learning algorithms to predict an individual's treatment response [27].

Areas of future development include using AI for drug discovery to control epileptiform discharges, which could be monitored using electroencephalography. There is a putative role of finding “repurposed” medications that can work for epilepsy, e.g., propranolol and candesartan have been found to work for migraine and these discoveries were serendipitous. Given a large

enough dataset, deep learning can alert clinicians to medications that decrease seizure frequency but that have not previously been considered for the treatment of epilepsy.

Machine Learning for Epilepsy Surgery

Barbour and colleagues reported another interesting use-case for machine learning by applying natural language processing to screen for risk factors in sudden unexplained death in epilepsy (SUDEP) so that the clinician can warn the patient and their family of what to look out for in order to reduce the risk of SUDEP occurring and to identify good surgical candidates. In their work, they reviewed 4000 manually coded patient charts and split the text into sentence tokens by the presence of periods, line breaks, three or more white spaces, and headings. A total of 579 (14%) were documented as having GTCS, 464 (12%) had refractory epilepsy, and 163 (4%) were potential or previous epilepsy surgery candidates. There was good interrater reliability for GTCS ($\kappa = 0.64$) and epilepsy surgery (0.72), and it was excellent for refractory epilepsy (0.89). Although varied, promising results were also reported here; for instance, their away test-set performance ranged from low to very high, including for GTCS (sensitivity = 0.94, PPV = 0.72, F-measure = 0.81), refractory epilepsy (0.86, 0.69, 0.76), and epilepsy surgery candidacy (0.74, 0.41, 0.53) [28].

The use of AI and machine learning in anesthetic depth recognition has its origins using EEG activity dated in the mid to late 1990s [29, 30]. Through evolved computational power, epilepsy for surgical anesthetic depth analysis and consciousness calculations using EEG have reached a stage where implementation is within the clinician's reach. Features that have been extracted are wide ranging from fractal and temporo-spectral characteristics to entropy measurements [30]. Others, like Mirsadeghi and colleagues, have employed dimensionality reduction approaches to characterize awake and

anesthetized states. Sun et al. extracted a total of 102 features from six electroencephalography (EEG) channels undergoing routine polysomnography with kappa statistical improvement of the classifier's output with increasing training data but with comparable results to a human scorer [31].

Patients whose seizures are refractory to anti-epileptics can be selected as surgical candidates.

An area of application for deep learning has been in connectomic disease risk factor discovery, surgical selection, and postoperative outcome prediction. Connectivity patterns have clinical phenotypic variations, which means that aberrant connectivity also varies across individuals and in epilepsy these variants occur in limbic and extra-limbic areas. Gleichgerrcht et al. applied deep learning on 50 patients with unilateral temporal lobe epilepsy (TLE) to classify them either as having persistent disabling seizures (SZ) or becoming seizure-free (SZF) at least 1 year after epilepsy surgery. They report a classification accuracy of their trained neural network to show a positive predictive value (PPV; seizure freedom) of $88 \pm 7\%$ and a mean negative predictive value (NPV; seizure refractoriness) of $79 \pm 8\%$ showing the potential of deep learning for postoperative outcome prediction [32].

Successful resection of the epileptogenic focus usually requires the ability to differentiate between normal brain regions and their epileptogenic foci that will enable the delineation of the normal-epileptogenic tissue boundary. The gold standard for monitoring this involves electrocortical stimulation and mapping (ESM) using an invasive grid electrode placement during the intraoperative period allowing functional localization of eloquent tissue. The problem here is that there is a risk of the surgeon provoking after-discharge seizures during the operative process, which consequently led to the introduction of the less risky alternative of electrocorticographic functional language mapping (ECoG-FLM). Deep learning using LSTMs (long short-term memory) has been applied on ECoG-FLM datasets to augment detection of the eloquent

zone with an 83.05% proposed accuracy and a sensitivity of 85%. Their data was from 637 electrode recordings, with 128 electrodes at 1200 samples per second generated millions of samples offer a decent training sample for training [33].

Machine Learning for Electrophysiological Migraine Detection

Migraine remains a highly prevalent and disabling neurological disorder, characterized by recurrent neuralgiform headaches. The diagnosis of migraine is still routinely biased toward clinical interviews, patient diaries, and physical examinations. There are currently no robust adjunct investigations to differentiate between various migraineous syndromes because there is equipoise as to whether there is a detectable electrophysiological substrate associated with migraine. Nevertheless, electrophysiological approaches, such as somatosensory evoked potentials (SSEPs), have shown potential to support identification of clinical fluctuations [34].

Machine learning has recently been applied to electrophysiological migraine detection [34]. The study recruited 42 migraine patients, including 29 interictal and 13 ictal, compared with 15 age-gender-matched healthy volunteers, and performed right median nerve somatosensory evoked potentials. Random forests, extreme gradient-boosting trees, support vector machines, K-nearest neighbors, multilayer perceptron, linear discriminant analysis, and logistic regression were applied to somatosensory evoked potential features in time and frequency domains. Accuracy ranges from 51.2% to 72.4% were achieved for the healthy volunteers-ictal-interictal. However, through appropriate feature selection, an improved accuracy of 89.7% for the healthy volunteers-ictal, 88.7% for healthy volunteers-interictal, 80.2% for ictal-interictal, and 73.3% for healthy volunteers-ictal-interictal classification tasks was achieved.

Other techniques such as therapeutic neuromodulation using electrophysiological transcranial magnetic stimulation are also making

an impact on migraine management. Researchers have also used time series-based algorithms like the line length algorithm used in responsive neurostimulation systems. Transcranial magnetic stimulation leverages intracranial electromagnetic currents to modulate neuronal activity; however its actual mechanism of action remains a black box mystery to us [35]. Applying deep learning to understand this “black box mystery” may open up the opportunity for useful discoveries. However, opponents would also rightfully argue that a deep learning approach is itself a black box which casts further unpredictability.

Deep Learning for Electromyography

Another neuro-electrophysiological method utilized in routine clinical practice for diagnostics is electromyography, which aids neuromuscular junction disease characterizations and records electrical activity produced by the skeletal muscle. Diseases that are easily diagnosed from EMG include the following:

- **Carpal tunnel syndrome (CTS):** A convolutional neural network achieved the best performance with AUC of 0.980, while support vector machines obtained an AUC of 0.943 in a testing dataset to diagnose CTC [36].
- **Cervical spondylosis:** One study was performed on 14 patients and 14 controls that underwent imaging of the cervical spine. Deep artificial neural network models were trained 1) to predict cervical spondylotic myelopathy (CSM) diagnosis and 2) to predict CSM severity. Model 1 was trained to predict a binary outcome and consisted of 6 inputs including 3 common imaging scales for the evaluation of cord compression, alongside 3 objective magnetic resonance imaging measurements. A reported model mean cross-validated accuracy was 86.50% (95% confidence interval, 85.16%–87.83%) with a median accuracy of 90.00% [37]. Wang et al. also applied a convolutional neural network-based multi-channel deep learning algorithm called EasiCSDeep on electromyography data, which

- consists of the feature extraction, spatial relationship representation, and classification [38].
- **Guillain-Barre syndrome:** One study used a dataset with 16 identified features to evaluate and perform a tenfold cross validation (10-FCV) using 15 single classifiers in two scenarios: four subtypes' classification and One versus All (OvA) classification. Their diverse classifiers ranged from decision trees, instance-based learners (k NN: k nearest neighbor), kernel-based learning (SVM, support vector machines), neural networks (SLP, MLP, and RBF-DDA), and rule induction learners (OneR, JRip) [39].
 - **Lou Gehrig's disease or amyotrophic lateral sclerosis (ALS):** Combination of three surface EMG markers achieved 90% diagnostic sensitivity and 100% diagnostic specificity, which were higher than solely using a single surface EMG marker [40].
 - Others include Lambert-Eaton syndrome, muscular dystrophy, myasthenia gravis, peripheral neuropathy, polymyositis, radial nerve dysfunction, and sciatica nerve dysfunction all of which will continue to benefit from the AI and machine learning revolution. In fact, Tapadar and George demonstrated a very practical machine learning implementation for myasthenia gravis diagnostics for starter clinicians interested in this area [41].

Electromyography can now be employed to facilitate gesture-based platforms for disabled patients using training datasets from able-bodied individuals. For example, one study using deep learning was used in the classification of these electromyography signals. Cote-Allard and colleagues reported the use of three different deep learning networks that used three different electromyographic modalities to train. Their inputs, (raw EMG, spectrograms, and continuous wavelet transform (CWT)) were trained on two datasets gathered from able-bodied individuals. Transfer learning methodology was used to enhance the performance for all three networks on two datasets. An offline accuracy of 68.98% for 18 gestures over 10 participants for the raw EMG-based ConvNet and 98.31% for 7 gestures

over 17 participants for the CWT-based convolutional neural network (ConvNet) was achieved [42].

Practical Machine Learning for EEG Analysis

Although texts may usually omit sections on how to find and curate datasets as well as where to start, we believe this to be important to facilitate knowledge dissemination and new discoveries for machine learning within the field. Various repositories have been created for those interested in practical approaches for EEG analysis using deep learning. Libraries available include Google's TensorFlow (<https://github.com/SuperBruceJia/EEG-DL>) which does have a steep learning curve if you have not programmed before.

Moreover, raw open-source EEG datasets can be downloaded for deep learning and other machine learning models, but data provenance and integrity must be verified. It will usually require the clinician partnering with an engineer or computer scientist to achieve this and to help develop the models. However, if one wants to brave it independently, we include a linked list of publicly available EEG datasets that can also be found at <https://github.com/meagmohit/EEG-Datasets>. Another resource that provides a starting point for most neurophysiological data visualizations can be found at <https://github.com/mne-tools/mne-python>. Tapadar and George also demonstrated a very practical machine learning implementation in the Python programming language for myasthenia gravis diagnostics for starter clinicians interested in this area discussed above [41].

If one is interested in the practical aspects, take note of the following:

1. Download Anaconda Python client, which offers you a visual interface and IPython framework, Jupyter Notebook, Spyder, etc.
2. Create a folder on your desktop and clone/download either of the above repositories EEG-DL (link and instructions provided above) into the folder where you intend to

- perform your electrophysiology or EEG analysis.
3. Follow the instructions on the EEG-DL link above, which are self-explanatory and involve interaction with either a Unix terminal or windows command line prompt.

Concluding Remarks

Highlighting the Tension Between Progress and “Model Explainability”

Although deep learning has considerable potential and we are seeing growing adoption and applications for clinical neurophysiology and EEG analysis, the lack of transparency about how models make decisions and the barriers to entry for AI education make it challenging to achieve acceptability among clinicians. For everyday life-changing decisions affecting patients, it is critical to see each step along the decision-making pathway as it could provide insights into the management of pathology and alert the medical establishment to patterns that have been missed. Clinicians are suspicious of and express concerns regarding black box decision-making as they hold both a duty of care and a duty of candor to patients. Consequently, there exists a clinical need to both identify what went wrong and be able to conduct a full risk and root cause analysis. Black box systems affect this diagnostic process. How then can one safeguard a patient from a clearly erroneous decision made by a neural network and where does the duty of candor lie? Questions have also been posed as to the ethical implications of who is responsible if the erroneous decision was made by an algorithm, as opposed to an error made by a human. These are subject areas that are being actively researched. In fact, Stephen Grossberg and others have proposed techniques such as adaptive resonance learning to enable model explainability [43]. In the coming years, we will be better equipped to understand these areas and appreciate model ingenuity, which will open the door to research and additional clinical discoveries. It is ironic that in our quest to understand the brain, we have made in-roads into the building of models of the brain that are now

transferable to solving other problems outside the brain. We are now modeling its function, behaviors, and pathology and designing extrinsic constructs to understanding but a small portion of its inherent frameworks, and yet we are still only scratching the surface to its potential. The evolution of AI promises to help us answer some of these ongoing puzzles of the brain, which will be beneficial for clinical neurophysiology as a whole.

References

1. Ghatol D, Widrich J. Intraoperative neurophysiological monitoring. [Updated 2020 Oct 12]. In: StatPearls [Internet] Treasure Island (FL): StatPearls Publishing; 2020.
2. Hinton G, Osindero S, Teh Y. A fast learning algorithm for deep belief nets. *Neural Comput.* 2006;18(7):1527–54.
3. Fisher R, Acevedo C, Arzimanoglou A, Bogacz A, Cross JH, Elger CE, et al. ILAE official report: a practical clinical definition of epilepsy. *Epilepsia.* 2014;55(4):475–82.
4. NICE. Overview | Epilepsies: diagnosis and management | Guidance 2020. <https://www.nice.org.uk/guidance/cg137>
5. Scheffer I, Berkovic S, Capovilla G, Connolly M, French J, Guilhoto L. ILAE classification of the epilepsies: position paper of the ILAE commission for classification and terminology. *Epilepsia.* 2017;58(4):512–21.
6. Bouma H, Labos C, Gore G, Wolfson C, Keezer M. The diagnostic accuracy of routine electroencephalography after a first unprovoked seizure. *Eur J Neurol.* 2016;23(3):455–63.
7. Dunn M, Breen DP, Davenport RJ, Gray AJ. Early management of adults with an uncomplicated first generalised seizure. *Emerg Med J.* 2005;22(4):237–42.
8. Jackson A, Teo L, Seneviratne U. Challenges in the first seizure clinic for adult patients with epilepsy. *Epileptic Disord.* 2016;18(3):305–14.
9. Palka D, Yogarajah M, Cock H, Mula M. Diagnoses and referral pattern at a first seizure clinic in London. *J Epileptol.* 2017;25(1–2):31–6.
10. Karayannidis NB, Tao G, Xiong Y, Sami A, Varughese B, Frost JD, et al. Computerized motion analysis of videotaped neonatal seizures of epileptic origin. *Epilepsia.* 2005;46:901–17.
11. Ogura Y, Hayashi H, Nakashima S, Soh Z, Shibanoki T, Shimatani K, et al. A neural network based infant monitoring system to facilitate diagnosis of epileptic seizures. In: 37th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2015.
12. Milosevic M, Van de Vel A, Bonroy B, Ceulemans B, Lagae L, Vanrumste B, et al. Automated detection of tonic-clonic seizures using 3-D accelerometry and

- surface electromyography in pediatric patients. *IEEE J Biomed Health Inform.* 2016;20:1333–41.
- 13. El Azami M, Hammers A, Jung J, Costes N, Bouet R, Lartizien C. Detection of lesions underlying intractable epilepsy on T1-weighted MRI as an outlier detection problem. *PLoS One.* 2016;11(9):e0161498.
 - 14. Jasper H. Report of the committee on methods of clinical examination in electroencephalography. *Electroencephalogr Clin Neurophysiol.* 1958;10 (1958):370–5.
 - 15. Gemein LAW, Schirmeister RT, Chrabaszcz P, et al. Machine-learning-based diagnostics of EEG pathology. *NeuroImage.* 2020;220:117021.
 - 16. Siuly S, Li Y, Zhang Y. *Electroencephalogram (EEG) and its background.* In: *EEG signal analysis and classification.* Cham: Health Information Science Springer; 2016.
 - 17. Lopez de Diego S. Automated interpretation of abnormal adult electroencephalography. Master's thesis, Temple University. 2017.
 - 18. Schirmeister RT, Gemein L, Eggensperger K, Hutter F, Ball T. Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum Brain Mapp.* 2017;38:5391–420.
 - 19. Roy S, Kiral-Kornek I, Harrer S. ChronoNet: a deep recurrent neural network for abnormal EEG identification. In: *Conference on artificial intelligence in medicine in Europe.* Springer; 2019.
 - 20. Obeid I, Picone J. The temple university hospital EEG data corpus. *Front Neurosci.* 2016;10:196.
 - 21. Amin S, Hossain M, Muhammad G, Alhussein M, Rahman M. Cognitive smart healthcare for pathology detection and monitoring. *IEEE Access.* 2019;7: 10745–53.
 - 22. Alhussein M, Muhammad G, Hossain M. EEG pathology detection based on deep learning. *IEEE Access.* 2019;7:27781–8.
 - 23. Van Leeuwen K, Sun H, Tabaeizadeh M, Struck A, Van Putten M, Westover M. Detecting abnormal electroencephalograms using deep convolutional networks. *Clin Neurophysiol.* 2019;130(1):77–84.
 - 24. Bjellvi J, Olsson I, Malmgren K, Wilbe Ramsay K. Epilepsy duration and seizure outcome in epilepsy surgery: a systematic review and meta-analysis. *Neurology.* 2019;93(2):e159–66.
 - 25. Bell G, Sinha S, Tisi JD, Stephani C, Scott C, Harkness W, et al. Premature mortality in refractory partial epilepsy: does surgical treatment make a difference? *J Neurol Neurosurg Psychiatry.* 2010;81(7):716–8.
 - 26. An S, Malhotra K, Dilley C, Han-Burgess E, Valdez JN, Robertson J, et al. Predicting drug-resistant epilepsy – a machine learning approach based on administrative claims data. *Epilepsy Behav.* 2018;89:118–25.
 - 27. Abbasi B, Goldenholz D. Machine learning applications in epilepsy. *Epilepsia.* 2019;60(10):2037–47.
 - 28. Barbour K, Hesdorffer D, Tian N, et al. Automated detection of sudden unexpected death in epilepsy risk factors in electronic medical records using natural language processing. *Epilepsia.* 2019;60:1209–20.
 - 29. Eagleman S, Drover D. Calculations of consciousness: electroencephalography analyses to determine anesthetic depth. *Curr Opin Anaesthesiol.* 2018;31(4):431–8.
 - 30. Shalbaf A, et al. Monitoring the depth of anesthesia using a new adaptive neuro-fuzzy system. *IEEE J Biomed Health Inform.* 2017;22:671–7.
 - 31. Sun H, et al. Large-scale automated sleep staging. *Sleep.* 2017;40(10):zsx139.
 - 32. Gleichgerrcht E, Munsell B, Bhatia S, et al. Deep learning applied to whole brain connectome to determine seizure control after epilepsy surgery Deep learning applied to whole brain connectome to determine seizure control after epilepsy surgery. *Epilepsia.* 2018;59:1643–54.
 - 33. RaviPrakash H, Korostenskaja M, Castillo EM, Lee KH, Salinas CM, Baumgartner J, et al. Deep learning provides exceptional accuracy to ECoG-based functional language mapping for epilepsy surgery. *Front Neurosci.* 2020;14:409.
 - 34. Zhu B, Coppola G, Shoaran M. Migraine classification using somatosensory evoked potentials. *Cephalalgia.* 2019;39(9):1143–55.
 - 35. Fatemi-Ardekani A. Transcranial magnetic stimulation: physics, electrophysiology, and applications. *Crit Rev Biomed Eng.* 2008;36(5–6):375–412.
 - 36. Ardakani AA, et al. Diagnosis of carpal tunnel syndrome: a comparative study of shear wave elastography, morphometry and artificial intelligence techniques. *Pattern Recogn Lett.* 2020;133:77–85.
 - 37. Hopkins B, Weber KA 2nd, Kesavabhotla K, Paliwal M, Cantrell DR, Smith ZA. Machine learning for the prediction of cervical spondylotic myelopathy: a post hoc pilot study of 28 participants. *World Neurosurg.* 2019;127:e436–42.
 - 38. Wang N, Cui L, Huang X, Xiang Y, Xiao J. EasiCSDeep: a deep learning model for cervical spondylosis identification using surface electromyography signal. *ArXiv.* 2018;abs/1812.04912.
 - 39. Canul-Reich J, et al. A predictive model for Guillain–Barré syndrome based on single learning algorithms. *Comput Math Methods Med.* 2017;2017:8424198.
 - 40. Zhang X, et al. Machine learning for supporting diagnosis of amyotrophic lateral sclerosis using surface electromyogram. *IEEE Trans Neural Syst Rehabil Eng.* 2014;22(1):96–103.
 - 41. Tapadar A, George AGA. Painless prognosis of myasthenia gravis using machine learning. Standford University; 2018.
 - 42. Cote-Allard U, Fall CL, Drouin A, Campeau-Lecours-A, Gosselin C, Glette K, Laviolette F, Gosselin B. Deep learning for electromyographic hand gesture signal classification using transfer learning. *IEEE Trans Neural Syst Rehabil Eng.* 2019;27(4):760–71.
 - 43. Grossberg S. Adaptive resonance theory: how a brain learns to consciously attend, learn, and recognize a changing world. *Neural Netw.* 2013;37:1–47.



Artificial Intelligence in Forensic Medicine

126

Thomas Lefèvre

Contents

Introduction	1768
Evidence and Individualization in Forensic Medicine	1769
Data, Information, and Evidence in Medicine and in Law: AI at Every Step in the Process	1770
Personalized Medicine, Individualized Sentences: A Shift from the Group to the Individual in Society	1770
Forensic Reasoning	1771
Artificial Intelligence and Assisted Decision-Making in Forensic Medicine	1771
AI and Clinical Forensic Medicine	1771
AI and Forensic Pathology	1772
AI and Medical Expertise in the Legal Context	1772
AI on the Borders of Forensic Medicine	1772
Artificial Intelligence for Forensic Medicine Research	1773
Potential Trends and Future Challenges	1773
Cross-References	1773
References	1773

Abstract

Forensic medicine lies at the crossroads between medicine and justice, with a particular connection to criminal law. This field can be

broken down into two major areas: clinical forensic medicine, which focuses on the living, and forensic pathology, which focuses on the dead. The use of data, and in particular artificial intelligence (AI), in this context faces two distinct challenges. First, there needs to be a discussion about the concept of evidence both upstream and downstream. A distinction must be made between scientific evidence, which is used to train algorithms and to support forensic reasoning in order to produce valid and robust information and legal evidence, which can include scientific data from investigations and

T. Lefèvre (✉)
IRIS Institut de Recherche Interdisciplinaire sur les enjeux Sociaux, UMR8156 CNRS – U997 Inserm – EHESS – Université Sorbonne Paris Nord, Paris, France

Department of Forensic and Social Medicine, AP-HP,
Jean Verdier Hospital, Bondy, France
e-mail: thomas.lefeuvre@univ-paris13.fr

interviews, as well as data produced by algorithms. The second challenge is individualization. One of the problems found in medicine, namely the application to a particular patient and situation of statistical knowledge established based on homogeneous groups whose study characteristics do not necessarily correspond to those of the patient, is only exacerbated in the field of forensic medicine. Not only must scientific knowledge be individualized, ensuring the validity of the algorithm to the specific case, without any learning or inclusion bias, but legal reasoning and the sentences handed down must be individualized too. AI can be used to support doctors' decision-making in forensic medicine, and it can also be used to structure research necessary for the evolution of scientific knowledge in forensic medicine. To date, the AIs used to support decision-making remain fairly immature.

Keywords

Forensic science · Forensic medicine · Legal medicine · Violence · Police custody · Artificial intelligence · Death · Expert · Forensic pathology · Autopsy

Introduction

Forensic medicine lies at the crossroads between medicine and justice, with a specific connection to criminal law [1]. The scope of this field and the prerogatives of medical examiners depend in part on each country's criminal law and on the respective roles of the actors involved: the role of the medical examiner in continental Western Europe is not the same as that of the coroner in English-speaking countries. Doctors, technical and scientific police, investigators, judges, and sometimes even anthropologists, dentists, and criminologists are just some of the many specialists called upon to help complete certain missions, organized according to the established procedure and the direction of the investigation. In this chapter, we will focus primarily on the medical dimension of forensic medicine and on the primary responsibilities of the medical

examiner. We will also discuss how this field intersects with others.

Forensic medicine can be broken down into two major areas, separated by death: clinical forensic medicine, which focuses on the living, and forensic pathology, which focuses on the dead [1]. Depending on the country, the first area may include:

- Crime scene forensics: estimating the functional impact on victims of intentional violence (interpersonal, physical, psychological, and sexual violence, as well as abuse) or unintentional violence (road traffic accidents), which may help to qualify what kind of offence has been committed or what kind of compensation victims are entitled to.
- Medical treatment for those held in custody: examining and treating those in prison or in police custody.
- Estimating the age of adolescent migrants: depending on the laws of each country, the legal provisions for migrants differ depending on whether the person is a minor or not.

The second area, forensic pathology, primarily includes:

- Determining the cause and mechanism of death, especially in cases of violent, suspicious, or unexpected death, which may have legal repercussions for a third party
- Initial observations and *in situ* sampling, in places where bodies or human remains are found
- Conducting external examinations: these may be carried out when removing a body or in special examination rooms. Investigating and making observations of any traumatic injuries and of any distinctive or identifying signs
- Conducting autopsies, including a macroscopic examination of the body and organs. This may also include further analyses, such as microscopic examinations (anatomical pathology), toxicological and genetic testing, etc.
- Postmortem identification: when a person's identity cannot be easily determined (due to decomposition, mutilation, only finding partial remains, etc.) or is unknown when the body is discovered

We will now review several opportunities for using AI in these different areas, without discussing them from too practical a point of view or in too much detail, since they will be covered in ► Chap. 127, “AI in Forensic Medicine for the Practicing Doctor.”

There are many complicated issues bound up in these two areas: health law, ethics, medical expertise (in the assessment of injuries and bodily harm for the compensation of victims), and psychiatric expertise (in the evaluation of the criminal responsibility of the accused and the assessment of a person’s dangerousness or risk of recidivism [2]).

There are two kinds of evidence that are central to forensic medicine: scientific evidence and legal evidence. As well as the question of evidence, there is the question of individualization: How can knowledge that has been established “generally” be applied with any certainty and suitability to specific cases or people? Artificial intelligence, much like P4 medicine (personalized, participatory, preventative, and predictive) [3], may be able to contribute significantly to answering these questions.

Before it can become accepted and used on an everyday basis by the various actors involved in forensic medicine, more research must be done on and with AI. On the one hand, we need a better understanding of what can be expected from AI, of how much it can be trusted in different usage contexts. On the other hand, research needs to be more rigorous, using more scattered and varied data. AI can be used in the collection, federation, and formatting of the data needed for research, and later for practice.

Ethical issues are often closely related to legal issues: in forensic medicine, the ethics of AI and of its practical use raise several specific problems.

Evidence and Individualization in Forensic Medicine

Since the rise of evidence-based medicine (EBM), medicine and clinical practice have been based on knowledge produced within a well-defined methodological framework and have referred to a well-established hierarchy of levels of evidence [4]. The classic case is the use of randomized controlled trials to try to define the causality between an

intervention and an observed event, for example, to help prove the effectiveness of a treatment for a given pathology. The methodological quality, reproducibility, and reproduction of these studies help to establish a range of arguments for or against the effectiveness of the treatment, that is, to determine its effectiveness. These trials are carried out under so-called experimental conditions, which do not reflect real life and are more similar to a laboratory experiment. Both EBM and randomized controlled trials have faced criticism, particularly for the fact that the patients included in these trials are themselves so highly “selected,” not only by the trial’s recruitment parameters, but also by the various inclusion and exclusion criteria, that they barely resemble the wider body of patients who will ultimately receive the treatment under trial [5]. Statistical methods are used to establish the reality of an observed difference between the treatment group and the control group and to account for sampling fluctuations due to chance or inter-individual variability. Patients are organized into homogeneous groups based on a certain number of criteria. This assumes the existence of an “average patient,” who is represented n times (where n is the number of patients included in the group), rather than n unique individuals. The results are therefore statistical and not individual, due to how they were produced. These two limitations make it more difficult to validly apply the results of a trial to a specific patient who comes into a doctor’s office. This is the individualization problem for which n of 1 trial has, for example, been suggested as a potential solution [6]. The same problem arises in prevention, when risk factors are used: these are established and statistically “true,” but they cannot be applied with certainty or be fully adapted to a specific individual.

Patient treatment requires scientifically established and “proven” knowledge, but it also requires the ability to individualize this knowledge, to apply it specifically to one person.

This is also true in forensic medicine and in justice: evidence must be arrived at through a set of arguments, and this evidence must be as specific as possible to the case in question, that is, individualized. When someone goes on trial, they are not judged as if they were an “average” person from a homogeneous group of “similar people.”

Data, Information, and Evidence in Medicine and in Law: AI at Every Step in the Process

Evidence is therefore the result of an entire production process, which ensures quality and reliability using a range of different criteria. That is what makes it so valuable. Evidence is established based on a larger mass of different pieces of information, which, when brought together, solidify into evidence [7]. Today more than ever, this information may come from data collection, processing, and analysis. There are some authors in both science and medicine who have even abandoned the term evidence-based medicine, preferring to talk about data-based or data-driven medicine or decision-making [8]. According to some extreme positions, the entire process of producing evidence by bringing to light causal relationships has run its course, and only data, seen as a “pure” way of observing the world, is necessary and sufficient, at least when analyzed by the correct algorithms. The process of establishing evidence would thus be backed up by the “objectivity” of data and its analysis by algorithms that are supposedly more reliable than humans [9]. Reasoning would be replaced by algorithms, and decision-making would be at least partially delegated to them. Of course, AI is not only used in the analysis of data; today, it is mostly used in the collection, federation, and extraction of data. This is the case in law, for example, where it is sometimes necessary to review entire bodies of case law around a specific issue [10]. AI may be used to identify rules for deciding cases or simply to provide lawyers or judges with summary documents. In such instances, the text analysis and data structuring are done using AI algorithms.

Personalized Medicine, Individualized Sentences: A Shift from the Group to the Individual in Society

By definition, the statistical methods used in randomized controlled trials produce statistical results. It is not easy to make the shift from these collective results to individual cases [6]. One of the promises, or rather one of the hopes placed in

AI, is that it will solve this problem of the individualization of knowledge. How can such information, based on evidence produced for homogeneous groups, but also on individual data on thousands or millions of people and the opinions of experts, be used to arrive at an understanding that will apply to a given individual? This is the idea behind personalized medicine [3–6], and it is what is expected in criminal law too, where people and their situations are judged on a unique basis: Based on the unique characteristics of this person and this situation, how can we reach a valid and similarly unique outcome? Some see AI as a tool for using these different data sources to reach a valid decision for the person in question. The assumption is that all possible combinations have already, in one way or another, been observed and recorded and that all AI needs to do is to find the right combination to determine the right decision [9]. Of course, the reality is more complicated than that [11]: algorithms operate by finding similarities, and they are also based on the creation of homogeneous groups – though these groups are certainly more numerous and granular than those used in traditional statistics or when extrapolating results from a linear regression. In all cases, one critical requirement is that data collection be as exhaustive as possible within the area of interest. This means that all of the relevant data about situations that may be encountered in forensic medicine (traumatic injuries, injury mechanisms, injury dates and timelines, psychological and physical functional impacts of violence on victims, etc.) and in the legal sphere (context of the offence, social context, etc.) would need to be identified, recorded, digitized, and perhaps even shared and standardized. Even at this point, we can begin to trace out two main regimes: Most offences are relatively minor (theft, minor acts of violence, breaking the speed limit, drink driving, etc.), so most of the data and situations encountered will concern these offences. Here, individualization is less critical, because the repercussions are less severe for the individual and for society. Only a minority of offences are major crimes such as homicide, rape, or terrorist acts. In these instances, it might be useful to have some kind of tool to support decision-making, which would help to individualize the decision while

remaining limited in its role and respectful of the collective rules inscribed in law. However, in such situations, AI is much less likely to seem reliable or relevant.

Forensic Reasoning

Over the last decade or so, forensic medicine has seen the publication of several studies, by researchers such as Biedermann and Taroni, that have attempted to establish a general framework for the individualization of results [12]. The framework used is Bayesian, in that it creates representations of knowledge based as much on data as on expert opinions or scientifically established data. Even the parameters can be estimated using the available data. These models make it possible to calculate conditional probabilities and to account for (quantify) the uncertainty related to an event or decision. This means that it is possible to calculate the probability that a certain DNA sample will match the DNA of a certain individual, given that certain characteristics, specific to the person and the context, have been observed. Of course, when specific combinations of characteristics are poorly represented in the data, the level of uncertainty will be higher. This approach could serve as a formal and well-established framework for the use of AI in courts, with the advantage that this kind of AI is more easily explained than other AI models, such as deep learning, or models with large numbers of parameters and hyperparameters, which are impossible to discuss easily with nonspecialists. Bayesian models are basically a graphic representation of the interdependencies between variables (directed graph), which can be used in explanations or to stimulate discussion. To our knowledge, however, this framework has not been used extensively in courts or by forensic doctors in their day-to-day practice. It also has its own shortcomings, such as problems with estimating a priori probabilities, which are central to the Bayesian framework, or probability interpretations (frequentist vs. subjectivist, for example) [13, 14].

Beyond individualization, AI can also be used to help with decision-making.

Artificial Intelligence and Assisted Decision-Making in Forensic Medicine

In this section, we will assume that the AIs in question use algorithmic methods known for their effectiveness in AI and machine learning. We will not discuss specific uses in detail, since they are the subject of their own separate chapter (see ► Chap. 127, “AI in Forensic Medicine for the Practicing Doctor”).

AI and Clinical Forensic Medicine

To date, there has been little, if any, research done on the possible uses of AI in clinical forensic medicine. There have been some multivariate models for estimating the ages of adolescent migrants [15], but they all share the same major flaw, namely that they were not designed based on the populations they are meant to assess [16]. The performance of these models is difficult to measure because the studies themselves do not follow the existing recommendations for AI model reporting (CHARMS or TRIPOD, for example [17]). Some authors have shown an interest in using AI to “diagnose” rape [18]. From an ethical and scientific standpoint, this use of AI should be viewed with caution, since rape is a legal classification and not a medical diagnosis (and therefore not a medical prerogative), and because there is no medical way to prove rape, only indicators that there has been sexual contact. In law, what characterizes rape is whether or not the act was consensual.

Some studies have been done or are currently under way on the possible use of AI to assess the functional impacts of violence on victims [19]. This would help doctors to quantify the duration of such impacts. They could also add the level of uncertainty of this estimate when communicating it to judges, knowing that different practices are employed in different medical centers and by different doctors. The acceptability and potential adoption of AI in this context have been studied using mixed (epidemiological and anthropological) methods, revealing significant fears among judges (fear of doctors being replaced by AI, then of being replaced themselves), but also

opportunities for them to integrate AI into their reasoning and to support their decisions [20]. A systematic review is currently under way, focusing on the role of AI in predicting PTSD among victims of sexual violence during their first consultation with a forensic medical professional [21]. While a handful of studies have already been published on AI and PTSD, none has yet focused on sexual violence specifically.

To date, there have been around ten studies published in clinical forensic medicine journals.

AI and Forensic Pathology

Here, we will not discuss all of the existing examples from forensic pathology, which are discussed more exhaustively in ► Chap. 127, “AI in Forensic Medicine for the Practicing Doctor.” We will, however, look at several representative examples: postmortem identification, estimation of time since death, and determining the cause and mechanism of death.

AI is used to help medical examiners identify bodies and remains by estimating several important characteristics of the body: age at death, sex, and morphology (height, weight, and body mass index). These AI draw most of their input data from medical imagery [22, 23].

Time since death estimates is generally based on observations made during external examinations of the body. These observations are not limited to clinical observations, but also include taking measurements (body temperature) and samples of various kinds, from sampling body fluids for biochemical analyses to sampling insects to collect forensic entomological information [24]. Photography is also used, for example, to estimate the time since death based on corneal clouding [25].

Determining the causes and mechanisms of death is a central part of thanatology and the primary goal of the autopsy. There are several AIs that can help medical examiners to determine the cause of death, which use natural language processing to consider different kinds of documents, such as autopsy reports, death certificates, or electronic health records (EHRs) [26]. These AIs also use postmortem imaging data, something

that is becoming more and more common with virtual autopsies (virtopsy [27]). They can detect internal hemorrhaging and potentially lethal injuries [28], as well as traces such as certain algae (diatoms) in body tissue to suggest – or rule out – the possibility of drowning [29].

These applications remain somewhat immature, with most arising from more or less completed research projects. They are rarely approved (external validation) for certified clinical use. The progress of these AIs is discussed in ► Chap. 127, “AI in Forensic Medicine for the Practicing Doctor.”

AI and Medical Expertise in the Legal Context

In the field of forensic expertise, AI is mostly used in relation to psychiatry. These approaches aim to improve or replace actuarial methods for evaluating psychiatric dangerousness or the risk of recidivism or suicide [2, 30]. They can also be used to help determine criminal responsibility if the accused cites impaired judgment, either at the time of the event or on an ongoing basis, perhaps due to some psychiatric pathology. They mostly use data from neuroimaging.

Generally speaking, any forensic situation that uses scales is likely to benefit from AI, as it appears to be a fairly natural extension of the use of scores and collective and individual data. This is already the case for calculating insurance estimates for personal injuries and damages [10], where the goal is to quantify both the damages and the corresponding compensation. These situations are encountered routinely, which may justify the use of AI.

AI on the Borders of Forensic Medicine

The direct practice of forensic medicine is supported by a range of other professions, techniques, and actions that may all be required in the forensic context. A nonexhaustive list of examples would include the growing interest in AI in ballistics (for identifying weapon or projectile types [31]); in fingerprint searching and comparison for

identification purposes [32]; in genetic analyses, especially for DNA identification based on different types of traces or samples [33]; and in the examination of potentially forged documents, with the knowledge that modern forgers often use AI themselves to create false documents [34].

Artificial Intelligence for Forensic Medicine Research

Well before it is used to support decision-making, AI can be used upstream to help structure research and the data needed to create new AIs for forensic medicine. As mentioned above, access to the necessary data is critical, as most data is divided between different information systems, which may also belong to different actors (e.g., several different hospitals, courts [35]). Even if the proper authorizations are obtained, most of the relevant databases are not interoperable, and it is often the case that some of the documents have not been digitized or are simple scans. Significant work must therefore be done to federate and analyze all of these different data sources to give them some kind of structure. One relevant example is the ORFéAD network, which connects 12 forensic medicine centers across France and federates the data from routinely produced medical documents, extracting the data and structuring a multicenter search database [20]. This database has already made it possible to create a family of algorithms to support decision-making when evaluating functional impacts on victims of violence [36]. This kind of approach is likely to catch on, in some variation, both because of the diversification of data sources and types and because of the cost involved in this kind of research, which requires ever-increasing volumes of data, but which faces increasing challenges in terms of securing the human resources to conduct this kind of research.

Potential Trends and Future Challenges

As in most medical fields, the development of AI and its routine use have yet to find their place in forensic medicine. This task is complicated by the

fact that the acceptance, relevance, and use of AI in forensic medicine is not just up to doctors, but also judges and, depending on the case, lawyers, victims, and those who have committed offences too [36]. AI can be used as part of the toolbox available to these actors to shore up the set of arguments that are ultimately used to establish evidence. Any such tool must be questioned, systematically investigated, and must prove its relevance and robustness in the face of doubt. These algorithms must also be unimpeachable from an ethical standpoint. For example, there can be no suspicion of gender or ethnic bias in these algorithms [37], or else they will be disqualified as nonapplicable. That is not to say that the algorithms used in a legal context must be somehow more “unimpeachable” than those used in other medical fields – all such algorithms should be similarly free of suspicion. In the legal context, however, algorithms are naturally more likely to be the target of criticism and doubt.

Until now, the primary role of AI in forensic medicine has been to support research methods and peripheral tasks: it is mostly used in technical matters, at the interface of medical practice. AI must overcome significant challenges if it is to have a more central, more specific role in medical practice. It is not only medical examiners who need convincing, but other nonmedical actors too, including judges.

Cross-References

- [AI in Forensic Medicine for the Practicing Doctor](#)

References

1. Payne-James JJ. Forensic medicine, history of. In: Encyclopedia of forensic and legal medicine, vol. 2. Amsterdam: Elsevier; 2015. p. 539–67. <https://doi.org/10.1016/B978-0-12-800034-2.00203-2>.
2. Tortora L, Meynen G, Bijlsma J, Tronci E, Ferracuti S. Neuroprediction and A.I. in forensic psychiatry and criminal justice: a neurolaw perspective. *Front Psychol*. 2020;11:220. <https://doi.org/10.3389/fpsyg.2020.00220>.
3. Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine. *Nat Rev Clin Oncol*. 2011;8:184–7.

4. Godlee F. Evidence based medicine: flawed system but still the best we've got. *BMJ*. 2014;348:g440.
5. Bujega G, Kumar A, Banerjee AK. Exclusion of elderly people from clinical research: a descriptive study of published reports. *BMJ*. 1997;315:1059.
6. Lillie EO, Patay B, Diamant J, et al. The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? *Pers Med*. 2011 Mar;8(2):161–73. <https://doi.org/10.2217/pme.11.7>.
7. Latour B, Woolgar S. Laboratory life: the social construction of scientific facts. Los Angeles: Sage; 1979.
8. Chorev M, Shpigelman L, Bak P, Yaeli A, Michael E, Goldschmidt Y. A data-driven decision-support tool for population health policies. *Stud Health Technol Inform*. 2017;245:332–6.
9. Anderson C. The end of theory: the data deluge makes the scientific method obsolete. *Wired*. 2008. <https://www.wired.com/2008/06/pb-theory>
10. <https://www.data.gouv.fr/fr/reuses/predictice/>
11. Pigliucci M. The end of theory in science? *EMBO Rep*. 2009;10(6):534. <https://doi.org/10.1038/embor.2009.111>.
12. Biedermann A, Bozza S, Taroni F. The decisionalization of individualization. *Forensic Sci Int*. 2016 Sep;266:29–38. <https://doi.org/10.1016/j.forsciint.2016.04.029>.
13. Biedermann A, Garbolino P, Taroni F. The subjectivist interpretation of probability and the problem of individualisation in forensic science. *Sci Justice*. 2013 Jun;53(2):192–200. <https://doi.org/10.1016/j.scijus.2013.01.003>.
14. Biedermann A, Taroni F, Garbolino P. Equal prior probabilities: can one do any better? *Forensic Sci Int*. 2007;172(2–3):85–93. <https://doi.org/10.1016/j.forsciint.2006.12.008>.
15. Lefèvre T, Chariot P, Chauvin P. Multivariate methods for the analysis of complex and big data in forensic sciences. Application to age estimation in living persons. *Forensic Sci Int*. 2016;266:581.e1–9. <https://doi.org/10.1016/j.forsciint.2016.05.014>.
16. Pruvost MO, Boraud C, Chariot P. Skeletal age determination in adolescents involved in judicial procedures: from evidence-based principles to medical practice. *J Med Ethics*. 2010;36(2):71–4. <https://doi.org/10.1136/jme.2009.031948>.
17. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, Reitsma JB, Collins GS. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med*. 2014;11(10):e1001744. <https://doi.org/10.1371/journal.pmed.1001744>.
18. Fernandes K, Cardoso JS, Astrup BS. A deep learning approach for the forensic evaluation of sexual assault. *Pattern Anal Applic*. 2018;21:629–40. <https://doi.org/10.1007/s10044-018-0694-3>.
19. Lefèvre T, Lepresle A, Chariot P. Detangling complex relationships in forensic data: principles and use of causal networks and their application to clinical forensic science. *Int J Legal Med*. 2015;129(5):1163–72. <https://doi.org/10.1007/s00414-015-1164-8>.
20. <https://orfead.org/en/orfead-forensic/>
21. Troussel V, Seyller M, Dang C, Chariot P, Lefèvre T. Prédire et dépister précocement un trouble de stress post-traumatique chez les victimes d'agressions sexuelles – potentiels de l'intelligence artificielle en consultation. 51ème congrès international de médecine légale. Dijon; 2019.
22. Mesejo P, Martos R, Ibáñez O, Novo J, Ortega M. A survey on artificial intelligence techniques for biomedical image analysis in skeleton-based forensic human identification. *Appl Sci*. 2020;10(14):4703. <https://doi.org/10.3390/app10144703>.
23. Anderson NE, Harenski KA, Harenski CL, Koenigs MR, Decety J, Calhoun VD, et al. Machine learning of brain gray matter differentiates sex in a large forensic sample. *Hum Brain Mapp*. 2019;40:1496–506. <https://doi.org/10.1002/hbm.24462>.
24. Moore HE, Butcher JB, Day CR, Drijfhout FP. Adult fly age estimations using cuticular hydrocarbons and artificial neural networks in forensically important Calliphoridae species. *For Sci Int*. 2017;280:233–44. <https://doi.org/10.1016/j.forsciint.2017.10.001>.
25. Cantürk I, Özilmaz L. A computational approach to estimate postmortem interval using opacity development of eye for human subjects. *Comput Biol Med*. 2018;98:93–9. <https://doi.org/10.1016/j.combiomed.2018.04.023>.
26. Duarte F, Martins B, Pinto CS, Silva MJ. Deep neural models for ICD-10 coding of death certificates and autopsy reports in free-text. *J Biomed Inform*. 2018;80:64–77. <https://doi.org/10.1016/j.jbi.2018.02.011>.
27. Dirnhofer R, Jackowski C, Vock P, Potter K, Thali MJ. VIRTOPSY: minimally invasive, imaging-guided virtual autopsy. *Radiographics*. 2006;26(5):1305–33. <https://doi.org/10.1148/rg.265065001>.
28. Ebert LC, Heimer J, Schweitzer W, Sieberth T, Leipner A, Thali M, et al. Automatic detection of hemorrhagic pericardial effusion on PMCT using deep learning – a feasibility study. *For Sci Med Pathol*. 2017;13:426–31. <https://doi.org/10.1007/s12024-017-9906-1>.
29. Zhou Y, Zhang J, Huang J, Deng K, Zhang J, Qin Z, et al. Digital whole-slide image analysis for automated diatom test in forensic cases of drowning using a convolutional neural network algorithm. *For Sci Int*. 2019;302:109922. <https://doi.org/10.1016/j.forsciint.2019.109922>.
30. Linthicum KP, Schafer KM, Ribeiro JD. Machine learning in suicide science: applications and ethics. *Behav Sci Law*. 2019;37:214–22. <https://doi.org/10.1002/bls.2392>.
31. Monash University. <https://www.monash.edu/news/articles/monash-designs-technology-to-map-bullet-trajectory>. (2019). Accessed 25 Nov 2020.
32. Neumann C, Evett IW, Skerrett J. Quantifying the weight of evidence from a fingerprint comparison: a new paradigm. *J R Stat Soc Ser A*. 2012;175:371–416.

33. Taroni F, Biedermann A, Vuille J, Morling N. Whose DNA is this? How relevant a question? (a note for forensic scientists). *Forensic Sci Int Genet.* 2013 Jul;7(4):467–70. <https://doi.org/10.1016/j.fsigen.2013.03.012>.
34. <https://shuftipro.com/blog/fighting-identity-fraud-with-ai-enabled-id-document-verification>
35. Lefèvre T. Big data in forensic science and medicine. *J Forensic Legal Med.* 2018;57:1–6. <https://doi.org/10.1016/j.jflm.2017.08.001>.
36. Guez S, Laugier V, Saas C, Lefèvre T. L'IA, le légiste et le magistrat: traitement médico-légal des violences interpersonnelles. In: Julia G, editor. Sciences et sens de l'intelligence artificielle, Thèmes et commentaires. Dalloz; 2020.
37. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 2019;366(6464):447–53. <https://doi.org/10.1126/science.aax2342>.



AI in Forensic Medicine for the Practicing Doctor

127

Laurent Tournois and Thomas Lefèvre

Contents

Introduction	1778
AI Applications for the Forensic Medicine Practice	1778
Part A: Description of AI Applications for the Forensic Medical Doctor	1778
Part B: Applicability and Usefulness of AI in Current and Future Practices in Forensic Medicine	1783
Conclusion	1785
References	1786

Abstract

Forensic medicine is a specialized medicine domain which aims at providing medical expertise in justice courts. In practice, forensic physicians or forensic pathologists can be requested by magistrates or police officers to

contribute solving an investigation, assessing and delivering forensically relevant elements requiring a thanatological or clinical expertise. Since the last decade, applications based on artificial intelligence have emerged in order to help forensic medical doctors in their daily practice. Most applications are developed from recent developments in AI, such as machine learning techniques.

This chapter proposes to describe such applications with a special attention to the current practices in forensic medicine. A use case from the forensic medical doctor's point of view is presented and analyzed for each expertise field. Finally, the applicability and usefulness of artificial intelligence in the routine tasks of forensic medical doctors are discussed in order to analyze the possible impacts in the practices of current and future forensic medicine.

Supplementary Information: The online version of this chapter (https://doi.org/10.1007/978-3-030-64573-1_221) contains supplementary material, which is available to authorized users.

L. Tournois (✉)
BioSilicium, Riom, France

UMR 8045 BABEL, University of Paris, Paris, France
e-mail: laurent.tournois@biosilicium.fr

T. Lefèvre
IRIS Institut de Recherche Interdisciplinaire sur les enjeux Sociaux, UMR8156 CNRS – U997 Inserm – EHESS – Université Sorbonne Paris Nord, Paris, France

Department of Forensic and Social Medicine, AP-HP, Jean Verdier Hospital, Bondy, France

Keywords

Artificial intelligence · Forensic medicine · Forensic medical doctor · Forensic physician · Forensic pathologist

Introduction

Forensic medicine is a specialized medicine domain which aims at providing medical expertise in justice courts. In practice, forensic medical doctors can be requested by magistrates or police officers to contribute solving an investigation, assessing and delivering forensically relevant elements requiring a thanatological or clinical expertise. In thanatology (as defined in Fig. 1), the main roles of the forensic pathologist are postmortem identification, the determination of the causes of death, and the estimation of the postmortem interval (PMI) or the time elapsed between the death of a person and the finding of the body or the remains. In clinical forensic medicine, forensic physicians mainly perform clinical assessments, carry out treatments on living persons such as people victim of crimes or detainees in police custody, and may give an expert opinion in justice institutions [1].

Since the last decade, applications based on artificial intelligence (AI) have emerged in order to help forensic medical doctors in their daily practice. Those applications are developed from public research and by private companies. Therefore, they may not necessarily be mentioned by any scientific paper or patent for business purposes. For this reason, it is hard to acquire exhaustive knowledge of the AI approaches or technologies in the forensic medicine practices. Hence, an exhaustive systematic review of those technologies is not feasible to date.

In consequence, the expertise fields of forensic medicine in which AI-based approaches are identified or identifiable are presented in this chapter. First, AI-based applications developed to support, augment, or replace the forensic medical doctor in routine tasks are described. The applications that could enhance or create new tasks in daily practice are also considered. However, some may not be

mature enough to date to be used in real conditions due to a lack of validation experiments. That is why the maturity level of the identified AI applications should be taken into account to visualize the future implications of the forensic medicine practice. To do so, this maturity level is estimated by using the technology readiness level (TRL) scale defined in [2]. Briefly stated, the initial TRL values range from 0 (the basic principles of the application are observed) to 9 (the application is validated and its final form is used in real conditions). In this chapter, the TRL scale is reduced to three values corresponding to three main stages of maturity of an application, as described in Table 1. The higher the TRL value, the mature the application is for a use in real conditions. It is thus possible to categorize the advances of AI in each expertise field of forensic medicine. Second, a special attention is paid to the applications with the highest or closest TRL value for a routine use by forensic medical doctors. Indeed, they are the most prone to be used in daily practice. Finally, the applicability and usefulness of AI in the daily tasks of forensic medical doctors are discussed in order to analyze the possible impacts of AI in the practices of current and future forensic medicine.

AI Applications for the Forensic Medicine Practice

Part A: Description of AI Applications for the Forensic Medical Doctor

In this part, AI applications are described by forensic medicine expertise field, as given in the first paragraph of the introduction. For each expertise field, a use case, real or elaborated for illustration purpose, is described from the point of view of the forensic medical doctor.

Thanatology

Postmortem Identification

Postmortem identification aims at finding elements of identification to associate a body or its remains to an identity. To do so, forensic

Forensics-related terms

Term	Definition
Thanatology	Scientific study of death and its associated practices. For instance, external examination of dead bodies and autopsy (or internal examination) are part of thanatology practices in forensic medicine
External examination	Non-intrusive examination of a body. The forensic medical doctor looks for the presence of traumatic wounds, the signs of death and of cadaveric processes such as putrefaction. The forensic pathologist may also take samples for toxicology or identification purposes.
Forensic entomology	Scientific study of arthropods in a forensic context, for example for rough time of death estimation
Diatoms	Unicellular algae composed of a silica shell visible under microscope. During vital submersion, diatoms penetrate into tissues, such as brain, liver or kidney, so that they may be used as elements in favour of a positive diagnosis of drowning.

Data/AI-related terms

Term	Definition
Machine learning	Field of AI that aims at giving computer systems the ability to draw inferences from data without being explicitly programmed.
Deep learning	Field of machine learning in which computer systems may use one or several variants of neural networks to draw inferences from data.
Neural network	Set of mathematical functions (artificial neurons) organized as a network. The layout of neurons and the links between them enable to perform complex tasks, such as object classification and detection.
Data reconciliation	Comparison of data to validate an action or an hypothesis. For instance, in post-mortem identification, data reconciliation consists in comparing post-mortem data against antemortem data to identify a body.

Fig. 1 Definition of terms related to forensics or data sciences

Table 1 Technology readiness level (TRL) values adapted to assess the maturity of AI-based applications in the practice of forensic medicine. The aTRL value represents a range of successive initial TRL values

Initial TRL range	Adapted TRL or aTRL value (used in this chapter)	Designation	Example
1–2	1	Formulation of the concept of the application	The concept of developing an application to estimate the age of an individual from wrist and hand radiograph is formulated
3–7	2	Application in research and development stages	One or several AI models are being developed and validated for a use by forensic medical doctors
8–9	3	Application used or almost ready to be used by forensic medical doctors in real conditions	Forensic medical doctors use the application to estimate the age of an individual at the request of justice authorities

pathologists currently use non-comparative or comparative approaches. The non-comparative approaches consist in inferring or determining individual characteristics to establish the identity of a subject, whereas the comparative approaches aim at comparing antemortem and postmortem data to search or confirm its identity [3].

For both approaches, several AI applications have already been developed. In non-comparative approaches, AI is mainly used to estimate personal characteristics such as age, gender [4], stature [5], height, weight, and body mass index [6].

Most of AI applications designed for that kind of approaches are designed to determine age or

gender of an individual from imaging techniques used in traditional forensic medicine. Indeed, age and gender are mainly determined from radiographs, and new developments take advantage of magnetic resonance imaging techniques for gender determination from brain features [7]. Some AI applications are also developed for facial reconstruction purposes from odontology data [8].

In comparative approaches, AI applications are used to label data or to perform data reconciliation between antemortem and postmortem data. To do so, applied AI research in forensic anthropology and odontology are currently carried out. A plethora of AI applications are also emerging in genetics to help the experts analyzing complex genetic profiles. However, this domain is not really relevant in this chapter since the samples of genetic material are mostly analyzed by genetic experts who are not necessarily and even rarely forensic medical doctors. That is why it was decided not to describe the applications of AI in genetics in this section.

In forensic anthropology, AI is applied for craniofacial superimposition, especially during the skull-face overlay stage, that is the superimposition of skull models to face photographs to identify an individual [9]. In forensic odontology, AI is mainly used to classify dental restorations [10] and dental cusps [11] or for features extraction [12] that are used as elements of comparison between antemortem and postmortem data.

To date, it is not currently possible to assert with certainty that all of those applications are used by forensic medical doctors in daily practice. Although most of applications are proofs of concept validated on real data (average aTRL 2 as shown in Table 2), few are actually reported

to be used in daily or occasional practice in forensic medicine. However, some are already used by forensic pathologists, especially during disaster victim identification (DVI) missions. Indeed, AI was used in the identification process of several disasters such as the MH17 plane crash in Ukraine in 2014, the air crash in Tripoli in 2010, and the tsunami event in Thailand in 2004 [13]. According to the 2018 INTERPOL DVI guide, DVI processes should include four steps defined as scene examination, postmortem data acquisition, antemortem data acquisition, and data reconciliation between antemortem and postmortem data [14]. In practice, AI is mainly applied in the latter. For example, the forensic pathologist may deal with hundreds of victims. In this case, one may imagine that identity documents, such as passports, are present on the disaster site. Therefore, facial recognition may be used to identify a victim from those passports. However, if the victim's face is too damaged, it becomes difficult to compare the victim's face with the photograph given on the passport. Therefore, the forensic pathologist may struggle to identify one victim by using traditional facial recognition methods.

Nevertheless, with the recent development of AI, the forensic pathologist could use an AI-based software that extracts relevant features from a face image. Then the software would perform data reconciliation (as defined in Fig. 1) from the extracted features against a database of identified faces and would output a match probability of the best candidate. The results of the matching process would be taken into account in the final decision handled by the forensic pathologist. The identification process would thus be faster than using

Table 2 Overview of AI application details found in the expertise domains of forensic medical doctors. The average aTRL represents the average of the technology readiness

value (TRL) based on the adapted TRL scale, as defined in Table 1, regarding the number of papers found/tools developed for an expertise domain

Expertise domain	Number of papers found/tools developed	Year of the oldest publication found	Average aTRL
Identification	>70	2012	2
PMI estimation	<20	2010	2
Determination of the causes of death	<30	2012	2
Clinical forensic medicine	<10	2018	2

traditional facial recognition methods, and the medical doctor can identify other victims that require other identification methods such as anthropology-based methods.

Some applications are also developed to anticipate future DVI cases. For instance, a software including AI components would be used in Japan for the automated analysis of panoramic X-ray radiographs in dental records to aid the identification of dead individuals [15]. Besides, since 2018, panoramic X-ray radiographs converted into a standardized format have already been stored within a database on a network shared by several medical institutions. It is worth mentioning that identifying individuals by AI in large-scale disaster from electronic records suggests that a relevant database of antemortem identification records exists and is accessible, all the data stored in the database are standardized to a given format for information sharing, and AI is able to use that database to retrieve relevant data for identification purposes.

Postmortem Interval (PMI) Estimation

PMI is mostly estimated by forensic medical doctors during the external examination of the body. Indeed the succession of features that appear on the body after the death enables to date approximately the death of a person. It is worth mentioning that external examination is not the only way used to estimate the PMI. Forensic medical doctors who are experts on forensic entomology and forensic biochemistry may use other skills to estimate the PMI with more accuracy than the methods used during classic external examination. In forensic entomology, PMI is estimated by identifying the successive waves of arthropods that appear after the death on the body and in the area where the body is found. In forensic biochemistry, the concentration of electrolytes in the eyeball (vitreous humour) may be used to estimate the PMI. Complementary methods based on body and ambient temperature measures may also be explored by using Henssge's nomogram [3].

In scientific literature, AI applications are appearing to estimate the PMI by using data from the external examination, as well as from

forensic entomology, forensic biochemistry observations, and temperature-based methods.

For instance, applied AI research is carried out to estimate the PMI from eye opacity features extracted from postmortem images [16]. In forensic entomology research, AI is mainly used to determine the age or the species of forensically relevant blowflies from spectrograms [17] or photographs [18], respectively. In forensic biochemistry, AI takes measured chemical compound concentrations as input data to estimate the PMI [19]. Recent papers suggest the use of AI to extract features from mass spectrometry profiles for PMI estimation [20]. However, this approach may not be used in daily practice since it is not validated on human cadavers to date.

Other papers also describe applications relying on AI to estimate the PMI from microbiome data [21].

Moreover, indirect approaches relying on AI are explored. For example, the authors in [22] have developed an AI model to determine the ambient temperature of a site in the past in order to calculate the accumulated degree days, an estimate of PMI, more accurately than the traditional methods.

However, none of the AI-based applications mentioned above are currently used in practice by forensic medical doctors. Nevertheless, it would be easy to imagine a use case in which a software analyzes an eye photograph to estimate PMI during the external examination of the body. In this case, the forensic pathologist would first take a picture of an eye during external examination. Then, the software would output the PMI estimate with a confidence interval corresponding to the accuracy of the AI model. Therefore, this application may avoid additional biochemical analysis for PMI estimation purposes and would enable to save the time spent on analyses processes and administrative procedures.

Determination of the Causes of Death

One of the main goals of forensic pathologists is to determine the causes of death from the observations reported during the autopsy (as defined in Fig. 1) and eventually from the results of additional analyses.

Several AI applications in this expertise field are based on textual data, such as autopsy reports, death certificates, and clinical bulletins [23]. Those applications aim mainly at determining the underlying cause of the death among the possible causes listed by the medical doctor.

Some research is also carried out to determine the cause of death from analytical data [24]. However, the resulting application is only limited to assess specific causes of death. Besides, most of AI applications designed to directly determine the cause of the death are specific to one or some causes of death.

Other AI applications may provide elements to determine the causes of death, such as the detection of hemorrhages [25], the identification of fatal injuries [26], or the detection of diatoms (as defined in Fig. 1) for the diagnosis of drowning [27], from postmortem imaging data.

To date, the diagnosis of drowning cases by AI is one of the most mature AI-based applications (aTRL 2; see Table 1) in research and development. Indeed, the presence of diatoms in specific tissues is a good indicator of drowning cases [28]. However, the search of diatoms in tissue samples may be a very time-consuming task since only one diatom may be present in one tissue sample. The comparison between diatoms from tissue and water samples may help confirm the location where the drowning occurred. Hence, the current AI approaches in this field mainly consist in diatom detection and/or classification.

Currently, AI is not reported to help forensic medical doctors in daily practice for the determination of the causes of death. However, one may imagine a use case in which a drowning case is suspected. In that case, the forensic pathologist would proceed to a diatom search on microscope slides of tissue and/or water samples. Instead of scanning manually a slide through a microscope, an AI application may be used to detect and classify automatically the diatoms present in a microscope whole-slide image. The forensic pathologist would then analyze the diatoms found to provide elements in favor or not of a positive diagnosis of drowning. To do so, the applications designed for those purpose should be able to identify at least the common diatom species in water samples as

well as tissue samples with high precision and accuracy. Those applications should also be able to learn to detect and classify diatom species for which they have never been trained with. Indeed, an AI algorithm is specific to the categories of data it was trained with. For example, if the algorithm was trained with 20 different diatoms, it would not be able to identify more than those 20 diatoms as is. Therefore, the AI algorithm should be flexible enough to enable the recognition of other diatom species. Hence, when encountering new species, the AI algorithm must be trained again with data including the new species to be able to recognize them in the future.

Clinical Forensic Medicine

In clinical forensic medicine, the forensic physician may be asked to perform clinical analyses on living persons and interpret the results to answer a given issue such as determining the age of a living individual [3] such as unaccompanied adolescent migrants or determining the mental state of a person or assess bodily injuries [1]. However, to the best of the authors' knowledge, AI is not currently used by medical doctor in clinical forensic medicine. Nevertheless, studies involving AI approaches are performed in clinical forensic medicine, especially in forensic psychiatry. In this field, AI applications are mainly designed to assess the risk or recidivism of an offender [29] or the risk of suicide of an individual [30]. Those studies are mostly based on neuroimaging data or electronic health reports. In both approaches, AI aims at extracting features used to assess a risk.

However, AI is currently not used by psychiatrists due to several reasons. First, the existing tools to assess the risk of violence are not widely adopted by forensic psychiatrists [29]. Therefore, an AI application may fail at the validation stages due to the use of nonrelevant data. Second, forensic physicians may have some reservations about using AI in forensic psychiatry due to law and ethical issues [31]. That is why forensic psychiatry is a field in which AI is currently hard to implement for a use in real cases.

Other AI applications are developed in order to help the forensic physician in the examination of individuals alleging to have been raped. One of

the most advanced AI applications (aTRL 2; see Table 1) is the automated classification of colposcopy images for forensic evaluation purposes of sexual assault [32]. Currently, this application is not developed enough to be used in daily practice for rape investigation. Nevertheless, it is relatively easy to imagine a use case from this application. For instance, the forensic physician could examine a victim through a digital colposcope associated to a computer which includes an AI-based application. Each time a photo is taken, the AI application would detect and point out wounds on the image. The forensic medical doctor would then decide if the highlighted elements are relevant enough in the diagnosis of sexual assault and proceed to first-line treatments. Therefore, AI could act as an assistant for the physician in this case. However, the application developed by [32] is not accurate enough to be used in real conditions.

Part B: Applicability and Usefulness of AI in Current and Future Practices in Forensic Medicine

Since the last decade, more and more AI applications in forensic medicine have been developed mainly, thanks to the developments in information technology, the increase of data volume, and computational power [33]. Currently, AI is mainly used for machine learning purposes, in other words the learning of patterns or representations in given data to answer a specific issue. To date, AI is able to learn from different data types, such as numerical values, images, reports, videos, and graphs. Hence, the diversity of data types an AI may learn from enables to explore AI-based approaches for many goals in forensic medicine, such as the detection of anomalies in postmortem computed tomography (PMCT) scans [34], for instance. However, AI is not able to model new purposes without human intervention. Therefore, AI applications are currently designed for specific tasks only. For example, an AI model only trained to determine the age of an individual will not be able to determine the gender of a person or use other data types than one or ones used during the

learning process, if the AI is not modified by a human. Besides, all the AI applications mentioned in this chapter are trained for a sole task.

Nevertheless, AI is useful in forensic medicine, especially for detection and classification tasks. For instance, as described in section “[Determination of the Causes of Death](#),” AI may be used to detect and classify automatically diatoms from whole-slide images. Therefore, the search of diatoms would be handled by an AI application, which would thus aid not to miss any diatom present on a slide. AI applications may also be designed in order to optimize time-consuming tasks, such as data reconciliation in DVI cases, as mentioned in section “[Postmortem Identification](#).”

Although many research studies deal with AI-based approaches for forensic medicine (see Table 2), the AI applications developed for DVI purposes seem to be the only ones which are reported in the literature of the current practices. From this observation, one may infer that AI is not used in daily practice by forensic medical doctors. However, the difference between the number of AI-related research papers in forensic medicine (see Table 2) and the number of applications used in daily practice by forensic medical doctors may be due to several reasons. First, not all AI applications designed from research projects are implemented in real and operational environments. Indeed, they may only contribute to improve the knowledge and the literature of applied AI in forensic medicine. Second, most of research studies in AI applied to forensic medicine are derived from recent developments in AI, such as deep learning approaches (as defined in Fig. 1). Therefore, due to a lack of validation studies, it may be too early to use the resulting AI applications in practice. For instance, PMCT imaging is a relatively recent technique implemented in forensic medicine practices; thus, AI models developed to process PMCT data often lack reliable data to be implemented in real environments. Third, medical doctors may be reluctant to the use of AI in forensic medicine.

Indeed, forensic medical doctors must be trained to use AI models and interpret their results. However, some AI applications, especially those relying on dense neural networks, are considered as “black box” models. In other words, the

rationale or the explainability of an AI model for a specific task is not known or not clear enough to figure out how the model interprets data. For instance, if an AI model based on dense neural network is trained in order to determine the height of an individual from anthropometric measures, it would be hard to understand how the model interprets the data to output the height of an individual. This is due to the architecture of the model. Indeed, dense neural networks are composed of several layers of individual operations also called neurons, such as the layers of neurons that process electric signals in the brain. A neuron of a layer is connected to all the neurons of the previous layer, except for the first layer which contains input data. With few neurons, it is relatively easy to understand the meaning of operations. For example, with two neurons, one may be dedicated to the detection of spine curvatures, another one to the correlation between height and the length of the spine. However, the more the neurons and layers, the more difficult the explanation since layers are densely connected. Some interpretation methods have already been studied [35]; however, it is currently hard to clarify the AI interpretation processes for models based on dense neural networks. Nevertheless, it is worth mentioning that research projects in explainable or interpretable AI are currently thriving, so that one may expect in the future that more explicit interpretation methods would emerge.

Forensic medical doctors may also be reluctant to use AI in daily practice due to bias-related issues. Indeed, current AI applications in healthcare medicine may correlate the ethnic group an individual belongs to with a given issue. For instance, AI might discriminate against Black people with regard to the additional care provision in complex cases [36]. Therefore, an AI model may be biased by unwanted correlations. Hence, justice decisions that would partially result from biased elements provided by an AI could be unfair due to rationales based on non-admissible elements, such as the race of the defendant. Moreover, AI models may also be biased by data. Indeed, when AI models are trained on too little data, they may be prone to overfitting. In other words, the model fits so well the data it is trained

on that it is unable to generalize on new data, which leads to overrate the performance of the model. Nevertheless, bias identification and removal in AI models are subject to active research studies. To be applied in real conditions, AI models should thus be transparent and fair in order to prevent any misinterpretation based on spurious correlations. That is the reason why validation studies are prominent to prevent such biases in the lifecycle of an AI application.

In addition, the use of AI in justice procedures raises other ethical issues such as liberty or privacy violation [31]. For example, in a fictitious case, if an AI algorithm erroneously points out a person with a high risk of violence, that person could be placed in police custody or in hospital to receive psychiatric care, which is a case of liberty violation. This fictitious example may remind some science fiction movies, like Minority Report, which explore that kind of speculative and preventive punishment. However, no current legal system can condemn an individual on the sole basis of an estimated risk, without a well-characterized offense. Privacy may also be affected since the data used to train AI should originate from real cases. Nevertheless, regulations like the General Data Protection Regulations or the Health Insurance Portability and Accountability Act are emerging to supervise the storing and sharing of such data.

Forensic medical doctors may also be worried about being replaced by AI in the future. Actually, current AI technologies do not allow replacing forensic examiners. Although AI models may provide correct and accurate suggestions, the final decision remains entrusted to the forensic medical doctor since models may make mistakes. Indeed, as mentioned above, AI models are mostly designed to answer a given issue. Even though data-driven algorithms may be able to handle several tasks, they are not able to understand the meaning of data. The interpretation of data by AI is only based on correlations, patterns, or explicit rules given by a human being. Therefore, forensic medical doctors have a more complete understanding than AI, so that they are more skilled to detect and explain outliers in data than AI. In addition, forensic medical doctors should

be vigilant about the use of AI to determine or assess elements that cannot be defined on a clinical or physical basis. For instance, in the use case presented in section “[Clinical Forensic Medicine](#),” the definition of rape relies on the notion of nonconsensual intercourse that cannot be assessed nor proven through an AI or other methods. Hence, using an AI algorithm that proposes to determine the consensual nature of an intercourse based on physical observations is not relevant since there is no certain causal relationship between both. Therefore, forensic medical doctors should first assess the relevancy of an AI algorithm before using it in daily practice.

As a consequence, AI models may currently be regarded more like an assistant for forensic medical doctors than fully skilled experts.

Moreover, AI might be applied to enhance forensic medical doctors in their practices instead of replacing them. Indeed, the applicability of AI in several domains of forensic sciences, like ballistics and forensic medicine, enables transdisciplinary approaches to address a given issue. For example, the determination of a bullet trajectory from postmortem data is currently a research project that may be used in the future by forensic pathologists [37]. In that case, the forensic pathologist may request an analysis of PMCT scans by AI to determine the vital parts a bullet may have damaged in order to investigate fatal damages caused by the bullet.

However, AI models should not be transferable from one application to another without expertise in both applications. For example, AI applications used for postmortem purposes may not be relevant for antemortem purposes. Indeed, missing elements of the body, decomposition processes or predation marks that could alter the tissues, are limiting factors in postmortem identification. Therefore, clinical and postmortem identification by AI would require either an AI model trained on postmortem and antemortem data or two AI models, one specific to postmortem data and the other one specific to antemortem data. As a consequence, forensic medical doctors should be trained to select the right AI model to use by taking into account the advantages and the limitations inherent to the model.

Nonetheless, a positive opinion from forensic medical doctors about the use of AI in forensic medicine does not mean a successful implementation of AI in daily practice. Indeed, the expertise of forensic medical doctors enables to produce conclusions and evidence elements that should be admissible by all the justice, especially magistrates and advocates [38]. Moreover, in forensic medicine, the medical doctor deals with unique cases that should be subject to unique and personalized justice decisions. Therefore, it is hard to rely on algorithms based on statistical groups. Hence, in justice court, it could be relatively easy to argue against the use of statistical AI by explaining that each case is unique and must be considered with all its specificities, which are not included in any statistical rationale. Some authors have proposed a Bayesian approach to individualize results in forensic medicine; however, this is still limited for a daily practice [39].

Therefore, the applicability of AI in forensic medicine depends on the adoption of AI by forensic medical doctors as well as magistrates and barristers.

Conclusion

Finally, few applications based on AI approaches are reported to be used by forensic medical doctors to date. In practice, the applications which are really used by medical doctors are mostly developed for identification purposes and applied in DVI cases. However, hundreds of research projects involving AI have been conducted in forensic medicine, and research studies are still in progress. Currently, most of those studies and projects from research and development departments of public institutions or private organisms show promising results. The performance of AI models in forensic medicine ranges from approximately 30% for the less accurate model to approximately 99% for the best one depending on the application and the forensic expertise field. Although most of current AI applications are more than 75% accurate, many are trained on too little data, so that such models are subject to overfitting and their performance is overrated.

Indeed, this overfitting phenomenon produces models which are not able to generalize to other data than those that were used during the development of the models. Therefore, such models cannot be applied in daily practice. Moreover, the implementation of AI in forensic medicine is not a trivial process. Indeed, the AI-based applications produced during research stages require further validation studies with more real data to be implemented in real and operational environments. Moreover, forensic medical doctors may be reluctant to use AI in daily practice due to ethical, professional, and AI-inherent issues, such as the presence of biases and the lack of elements of AI interpretability in specific kinds of application.

Nevertheless, AI is a thriving field in which techniques are constantly improved, so that the potential of AI in forensic medicine purposes is not fully exploited to date. Therefore, AI applications in forensic medicine are promising data-driven tools that could enhance the forensic medical doctor in daily tasks.

References

1. Payne-James JJ. Forensic medicine, history of. In: Encyclopedia of forensic and legal medicine, vol. 2. Elsevier Science; 2015. p. 539–67. <https://doi.org/10.1016/B978-0-12-800034-2.00203-2>.
2. Mankins J. Technology readiness levels. A white paper. NASA; 1995.
3. Beauthier JP. Traité de médecine légale. 2nd ed. Brussels: De Boeck; 2011.
4. Mesejo P, Martos R, Ibáñez O, Novo J, Ortega M. A survey on artificial intelligence techniques for biomedical image analysis in skeleton-based forensic human identification. *Appl Sci*. 2020;10(14):4703. <https://doi.org/10.3390/app10144703>.
5. Czibula G, Ionescu VS, Miholca DL, Mircea IG. Machine learning-based approaches for predicting stature from archaeological skeletal remains using long bone lengths. *J Archeol Sci*. 2016;69:85–99. <https://doi.org/10.1016/j.jas.2016.04.004>.
6. Dantcheva A, Bremond F, Bilinski P. Show me your face and I will tell you your height, weight and body mass index. *Int Conf Pattern Recognit*. 2018;24:1–6.
7. Anderson NE, Harenski KA, Harenski CL, Koenigs MR, Decety J, Calhoun VD, et al. Machine learning of brain gray matter differentiates sex in a large forensic sample. *Hum Brain Mapp*. 2019;40:1496–506. <https://doi.org/10.1002/hbm.24462>.
8. Niño-Sandoval TC, Pérez SVG, González FA, Jaque RA, Infante-Contreras C. Use of automated learning techniques for prediction mandibular morphology in skeletal class I, II and III. *For Sci Int*. 2017;281:187. e1–7. <https://doi.org/10.1016/j.forsciint.2017.10.004>.
9. Damas S, Cordón O, Ibáñez O. Handbook on craniofacial superimposition – The MEPROCS Project. Springer; 2020.
10. Abdalla-Aslan R, Yeshua T, Kabla D, Leichter I, Nadler C. An artificial intelligence system using machine learning for automatic detection and classification of dental restorations in panoramic radiography. *Oral Surg Oral Med Oral Pathol Oral Radiol*. 2020;130(5):593–602. <https://doi.org/10.1016/j.oooo.2020.05.012>.
11. Raith S, Per Vogel E, Anees N, Keul C, Güth JF, Edelhoff D, et al. Artificial Neural Networks as a powerful numerical tool to classify specific features of tooth based on 3D scan data. *Comput Biol Med*. 2017;80:65–76. <https://doi.org/10.1016/j.combiomed.2016.11.013>.
12. Fan F, Ke W, Wu W, Xuemei T, Lyu T, Liu Y, et al. Automatic human identification from panoramic dental radiographs using the convolutional neural network. *For Sci Int*. 2020;314:110416. <https://doi.org/10.1016/j.forsciint.2020.110416>.
13. Plass Data. DVI System International. <https://www.plassdata.com/products-services/software-products.html#dvi>. Accessed 2 Dec 2020.
14. INTERPOL. 2020. <https://www.interpol.int/How-we-work/Forensics/Disaster-Victim-Identification-DVI>. Accessed 25 Nov 2020.
15. Takano H, Momota Y, Ozaki T, Shiozawa S, Terada K. Personal identification from dental findings using AI and image analysis against great disaster in Japan. *J Forensic Leg Investig Sci*. 2019;5:041. <https://doi.org/10.24966/FLIS-733X/100041>.
16. Cantürk I, Özilmaz L. A computational approach to estimate postmortem interval using opacity development of eye for human subjects. *Comput Biol Med*. 2018;98:93–9. <https://doi.org/10.1016/j.combiomed.2018.04.023>.
17. Moore HE, Butcher JB, Day CR, Drijfhout FP. Adult fly age estimations using cuticular hydrocarbons and Artificial Neural Networks in forensically important Calliphoridae species. *For Sci Int*. 2017;280:233–44. <https://doi.org/10.1016/j.forsciint.2017.10.001>.
18. Luquin MFH, Santacruz EV, Morales RAL, Vázquez CN, Zúñiga MG. Development of intelligence tools for recognizing cockroaches in the forensic entomology context. *Intell Syst*. 2017. <https://doi.org/10.1109/IntelliSys.2017.8324269>.
19. Muñoz-Barús JI, Rodríguez-Calvo MS, Suárez-Peñaanda JM, Vieira DN, Cadarso-Suárez C, Febrero-Bande M. PMICALC: an R code-based software for estimating post-mortem interval (PMI) compatible with Windows, Mac and Linux operating systems. *For Sci Int*. 2010;194:49–52. <https://doi.org/10.1016/j.forsciint.2009.10.006>.

20. Zhang J, Wei X, Huang J, Lin H, Deng K, Li Z, et al. Attenuated total reflectance Fourier transform infrared (ATR-FTIR) spectral prediction of postmortem interval from vitreous humor samples. *Anal Bioanal Chem.* 2018;410:7611–20. <https://doi.org/10.1007/s00216-018-1367-1>.
21. Johnson HR, Trinidad DD, Guzman S, Khan Z, Parziale V, DeBruyn JM, et al. A machine learning approach for using the postmortem skin microbiome to estimate the postmortem interval. *PLoS One.* 2016;11(12):e0167370. <https://doi.org/10.1371/journal.pone.0167370>.
22. Jeong SJ, Park SH, Park JE, Park SH, Moon T, Shin SE, et al. Extended model for estimation of ambient temperature for postmortem interval (PMI) in Korea. *For Sci Int.* 2020;309:110196. <https://doi.org/10.1016/j.forsciint.2020.110196>.
23. Duarte F, Martins B, Pinto CS, Silva MJ. Deep neural models for ICD-10 coding of death certificates and autopsy reports in free-text. *J Biomed Inform.* 2018;80:64–77. <https://doi.org/10.1016/j.jbi.2018.02.011>.
24. Lin H, Luo Y, Sun Q, Deng K, Chen Y, Wang Z, et al. Determination of causes of death via spectrochemical analysis of forensic autopsies-based pulmonary edema fluid samples with deep learning algorithm. *J Biophotonics.* 2020;13:e201960144. <https://doi.org/10.1002/jbio.201960144>.
25. Ebert LC, Heimer J, Schweitzer W, Sieberth T, Leipner A, Thali M, et al. Automatic detection of hemorrhagic pericardial effusion on PMCT using deep learning – a feasibility study. *For Sci Med Pathol.* 2017;13:426–31. <https://doi.org/10.1007/s12024-017-9906-1>.
26. Garland J, Ondruschka B, Stables S, Morrow P, Kesha K, Glenn C, et al. Identifying fatal head injuries on postmortem computed tomography using convolutional neural network/deep learning: a feasibility study. *J For Sci.* 2020;65(6):2019–22. <https://doi.org/10.1111/1556-4029.14502>.
27. Zhou Y, Zhang J, Huang J, Deng K, Zhang J, Qin Z, et al. Digital whole-slide image analysis for automated diatom test in forensic cases of drowning using a convolutional neural network algorithm. *For Sci Int.* 2019;302:109922. <https://doi.org/10.1016/j.forsciint.2019.109922>.
28. Farrugia A, Ludes B. Diagnostic of drowning in forensic medicine. In: Duarte NV, editor. From old problems to new challenges. InTechOpen; 2011. <https://doi.org/10.5772/19234>.
29. Cockerill RG. Ethics implications of the use of artificial intelligence in violence risk assessment. *J Am Acad Psychiatry Law.* 2020;48:345–9. <https://doi.org/10.29158/JAAPL.003940-20>.
30. Linthicum KP, Schafer KM, Ribeiro JD. Machine learning in suicide science: applications and ethics. *Behav Sci Law.* 2019;37:214–22. <https://doi.org/10.1002/bls.2392>.
31. Tortora L, Meynen G, Bijlsma J, Tronci E, Ferracuti S. Neuroprediction and A.I. in forensics psychiatry and criminal justice: a neurolaw perspective. *Front Psychol.* 2020;11:220. <https://doi.org/10.3389/fpsyg.2020.00220>.
32. Fernandes K, Cardoso JS, Astrup BS. A deep learning approach for the forensic evaluation of sexual assault. *Pattern Anal Appl.* 2018;21:629–40. <https://doi.org/10.1007/s10044-018-0694-3>.
33. Lefèvre T. Big data in forensic science and medicine. *J Forensic Leg Med.* 2018;57:1–6. <https://doi.org/10.1016/j.jflm.2017.08.001>.
34. Dobay A, Ford J, Decker S, Ampanozi G, Franckenberg S, Affolter R, et al. Potential use of deep learning techniques for postmortem imaging. *Forensic Sci Med Pathol.* 2020;16:671–9. <https://doi.org/10.1007/s12024-020-00307-3>.
35. Du M, Liu N, Hu X. Techniques for interpretable machine learning. *Commun ACM.* 2020;63(1):68–77. <https://doi.org/10.1145/3359786>.
36. Ledford H. Millions affected by racial bias health-care algorithm. *Nature.* 2019;574:608–9.
37. Monash University. 2019. <https://www.monash.edu/news/articles/monash-designs-technology-to-map-bullet-trajectory>. Accessed 25 Nov 2020.
38. Guez S, Laugier V, Saas C, Lefèvre T. L'IA, le légiste et le magistrat: traitement médico-légal des violences interpersonnelles. In: Julia G, editor. Sciences et sens de l'intelligence artificielle, Thèmes et commentaires. Dalloz; 2020.
39. Taroni F, Biedermann A. Probability and inference in forensic science. In: Buinsma G, Weisburd D, editors. Encyclopedia of criminology and criminal justice. New York: Springer; 2014. https://doi.org/10.1007/978-1-4614-5690-2_146.



Artificial Intelligence for Physiotherapy and Rehabilitation 128

Joseph Davids, Niklas Lidströmer, and Hutan Ashrafiān

Contents

Introduction	1790
AI in Physiotherapy and Physical Rehabilitation of Patients	1791
AI in Exergames and Serious Games for Early- Stage Physical Rehabilitation ..	1793
AI in Physiotherapy Education and Use of Simulation for Educating Physiotherapists	1799
AI and Physiotherapy Education	1800
AI for Robotic Assisted Physiotherapy	1800
AI for Physio-assisted Activity of Daily Living Monitoring	1800

J. Davids (✉)
Imperial College London NHS Trust, London, UK

Hamlyn Centre for Robotics and Artificial Intelligence,
Department of Surgery and Cancer, Imperial College
London, London, UK

National Hospital for Neurology and Neurosurgery Queen
Square, London, UK
e-mail: j.davids@imperial.ac.uk

N. Lidströmer
Department of Women's and Children's Health, Karolinska
Institutet, Stockholm, Sweden
e-mail: niklas.lidstromer@ki.se; niklas@lidstromer.com

H. Ashrafiān
Department of Surgery and Cancer, Imperial College
London NHS Trust, London, UK

Institute of Global Health Innovation, Imperial College
London, London, UK

Hamlyn Centre for Robotics and Artificial Intelligence,
Department of Surgery and Cancer, Imperial College
London, London, UK
e-mail: h.ashrafiān@imperial.ac.uk

AI and Virtual Reality for Physiotherapy and Rehabilitation	1801
AI for Physiotherapy-assisted Sensory and Balance Training	1801
AI for Assisted Wheelchair Users and Assisted Mobility Support	1801
AI for Inattention and Hemi-neglect Training	1802
AI for Respiratory Physiotherapy Management	1802
AI for Community Physiotherapy and Care	1803
AI for Cognitive Impaired Patients Needing Physiotherapy and Rehabilitation	1803
AI for Functional and Feedback Systems in Physiotherapy	1803
AI in Smart Watches and Wearables for Physiotherapy	1804
Future of AI in Physiotherapy	1804
References	1805

Abstract

Physiotherapy is a natural component of modern clinical medicine, and frequently the most efficient remedy to a wide range of medical conditions. Physiotherapists are an integral part of the medical team of professionals. After surgeries or accidents, especially those involving bones and joints, and against osteoarthritis and other conditions involving pain, the physiotherapeutic treatment is commonly used in clinical practice. It is often the sole curing ingredient. At other times patients usually present to physiotherapists with backpain that have sinister underlying pathologies such as cancer or cauda equina syndromes. Therefore, this chapter aims to delve into how this remedy could be augmented with artificial intelligence (AI). It will show how AI can increase the supportive frameworks of physiotherapy in the era of digitizing medicine, through a plethora of new applications, ranging from in real time video instructions in combination with pose and joint angle detections to give optimal feedback, to motivational and psychotherapeutic components to increase exercise precision and discipline. AI will contribute to making physiotherapy even more personalized, enduring, and further integrate cognitive behavioral therapy and virtual reality into the precise treatment regimens. It will also increase the frequency of exercise, since the patient can work out in several places, with the use of telemedicine for physiotherapy too.

Introduction

Physiotherapy is a specialty that involves guiding patients who have been affected by a pathological process, such as a neurological or traumatic insult leading to a disability, to help restore their function. It encompasses a group of skilled individuals working in tandem with a multidisciplinary team of occupational therapists, speech and language therapists, cognitive behavioral therapists, neuromuscular physiologists, specialist rehabilitation nurses, etc. Physiotherapists utilize integrative approaches adopting a holistic psycho-physico-biosocial approach to achieve this goal by allowing the patient to either directly regain their level of independence or reach a desired level of reduced carer dependence and improve their quality of life. The functional restoration is necessary to also limit the risk of further injury or relapse. It requires the right psychological, emotional, and physical support as well as direct and (as we will subsequently cover) indirect physical input to support the patient's needs. Patients and their carers usually require a significant amount of funding for (often long-term) packages of care post-discharge from the hospital to support their rehabilitation and subsequent recovery. There is also a need for an optimal interface between community physiotherapy and hospital physiotherapy to facilitate continuity of care, and to support quality improvement and feedback systems that

would later benefit patients going through a similar predicament. Additionally, an auditable trail and follow-up of patients that are discharged is also required to subsequently monitor progress and rehabilitation compliance. This all requires a significant proportion of funding and healthcare investment and is a socio-economic challenge for healthcare providers and policy makers.

These aforementioned areas are increasingly gaining recognition as key to a patient's recovery and are where machine learning can serve as significant adjuncts to physiotherapy service provision. Machine learning has been extensively covered in this book and so its introduction will not be heavily labored on here. However, as a sub-facet of artificial intelligence, machine learning algorithms are mathematical instructions allowing a computer or any electronic device to be programmed and learn to independently carry out an automated objective without the presence of human supervision. It has been applied to multiple fields with tremendous success and is touted to also impact physiotherapy. The extent of this impact continues to be a topic of considerable debate across multiple fields. Several chapters in this book have covered some of these applications and challenges as well as approaches to optimal machine learning methodology. A summary of the machine learning pipeline is presented in Fig. 1.

In physiotherapy, the re-training and rehabilitation of patients could benefit from artificially intelligent robotics or wearable and smart apps, and simulation for workforce education [1–10]. Indeed, progress has been made over the past decade where artificial intelligence has entered all aspects of healthcare delivery. However, ML also promises to help augment, but not yet replace the physiotherapist, whose role remains paramount to the pre and post-discharge recovery of a patients' cognitive and physical function.

AI in Physiotherapy and Physical Rehabilitation of Patients

This section discusses the use of artificial intelligence in physical therapy of patients. Physiotherapy for physical patient training is required for those who have had diseases that affect multi-

organ systems or undergone procedures to treat them. These diseases include [1, 3–6, 9–13]:

1. Rheumatological diseases, sports injuries resulting in back, neck and shoulder pain.
2. Diseases of the central nervous system such as age-related, developmental and movement disorders associated with Parkinson's disease, multiple sclerosis, Alzheimer's disease, amyotrophic lateral sclerosis, Huntington's disease, stroke syndromes, hydrocephalus, and brain tumors affecting the motor homunculus, etc. [11, 12].
3. Diseases of the cardiovascular system such as myocardial infarction and peripheral vascular disease necessitating rehabilitation [14].
4. Respiratory diseases such as infective exacerbations of chronic obstructive pulmonary disease, acute respiratory distress syndromes requiring intensive care support, and cystic fibrosis, which usually also requires breathing retraining to optimize tissue oxygenation and perfusion [15, 16].

Physical rehabilitation in these circumstances is usually performed by experienced specialist physiotherapists. About 50% of stroke survivors develop disabilities of motor function, which requires continuous rehabilitation including gait-based retraining [3, 12]. In this regard, machine learning algorithms have been developed for stroke recovery aiming to predict the scope of recovery based upon features such as duration of in-hospital stay, duration of stroke, and patients' Bartels index score [9, 12]. The intention is to enable a functional recovery estimate to be identified [12].

Another area utilizing AI for physical rehabilitation is the use of the OpenPose platform for posture detection, hand gesture detection and in sports physiotherapy for predicting how an individual can handle and throw a basketball [5]. It offers a cost-effective alternative to other methods that are resource-intensive and require expensive equipment. The OpenPose platform can be incorporated into a mobile app using an individual's phone camera to take photos and send images to the algorithm for automated posture assessment and classification. This becomes useful in patients

rehabilitating from lesions of the brain and spine that affect posture and sensori-motor function as is evident in brain and spinal tumours requiring surgical excision.

Since the physical side of rehabilitation is usually not completely effective without the psychological rehabilitation component as well, an additional extension to platforms like OpenPose system would be the use of natural language processing algorithms to build a chatbot that can then guide the patient through rehabilitation and pose training [4]. As such, chatbot systems can also be designed to offer further psychological and cognitive support to patients looking for additional training time after their physio sessions are completed for the day to augment healing, maintain routines, and stimulate compliance. Chatbots can also offer remote physiotherapeutic support by referring patients who require out-of-hours physiotherapy input within the community so that distressed patients can better engage with their physiotherapist and referred for the right service. It could also be applied to help follow-up patients for needs assessment, track their progress with their prescribed rehabilitation plan or determine whether there is non-compliance.

The above platforms facilitate a scenario where considerable amounts of information are captured to gain a complete picture of the patient's continuous psychological and emotional needs, which could in turn affect a patient's physical recovery. The need to engage other teams to order the right equipment and provide the care plans that the patient needs to rehabilitate is centered on the data available to the physiotherapist. Discussion with physiotherapists revealed that a considerable amount of time is spent processing this data and identifying missing data often from different disjointed sources. Most of these data sources are not electronic and even when they are, these systems poorly communicate with one another, making effective decision support challenging. As the data for patients significantly grows in scope and volume, especially for chronically ill patients, machine learning algorithms become necessary to derive insights that would aid the physiotherapists' decision making, assessment of patients for the optimal physical therapy regimen, etc. These

insights could support the selection of the right equipment needed for the patient to physically rehabilitate, the right companies with the required tools and the creation of advanced care plans to facilitate physical rehabilitation.

Microsoft invented the Kinect system for motion tracking, and it became popular for hands-free gaming. Kinect-based physical rehabilitation is another area demonstrating significant potential to change patient lives and physical needs due to its level of flexibility and low cost for human action capture [1, 3, 6, 13, 17]. It is marker-less which means that the patient does not need to be attached to wires. It also offers patient continuity from hospital rehabilitation into the home environment. There is also scope for physical feedback monitoring and a platform of external mentorship and interactivity with friends or others. In a study, the authors acquired multiple features from the patient including their skeleton joints and depth map features to assemble multiple classifiers. The areas that were analyzed and example of exercises that were developed to support patients included [3]:

- Range of motion exercises, which help patients with the mobility of their joints.
- Muscle strengthening exercises to gain better balance, mobility and ability to enjoy a normal lifestyle include squats, arm rowing exercises and modified push-ups, shoulder presses, etc.
- Balance exercises enable stroke and Parkinson's disease patients to improve their balance, or those with mobility-limiting muscle weakness to build strength and improve their balance.
- Flexibility exercises to support patients with upper motor neuron-related contractures associated with stroke and neurological disease, those experiencing back pain or those with Parkinson's disease as well as those with herniated disc conditions. Others also benefit from chest, hamstring, back and shoulder stretches. Although the degree of muscle stretch is difficult to objectively quantify and would benefit from AI-augmented measuring methods.
- Following surgery, physical rehabilitation is usually necessary to reduce post-operative pain and muscle stiffness, and to improve the patient's conditioning or to prepare them for

discharge to the community, where potential physical rehabilitation continues until full recovery.

AI in Exergames and Serious Games for Early- Stage Physical Rehabilitation

The process involves the patient performing rehabilitation exercises, such as stretching, captured by the Kinect sensor, which detects the person in the frame, analyzes the pose, recognizes the movement, and classifies the human action, the degree of difficulty and the activity performed [3, 18]. A reward-based system can be implemented to provide a psychological feedback incentive to engage patients.

Algorithms that have been discussed in the literature for human detection include Histogram of Gradient for low level feature identifier, Wavelet and Principal Component Analysis Shift, together with shape contexts [3, 19–24]. To improve detection speed, other authors have also proposed rejection cascades, but this can be computationally intensive and other low-cost algorithms have been proposed [3, 23].

Other algorithms include scale invariant feature transforms (SIFT) which are highly accurate with low computation time and have been used for hand-gesture-recognition [3, 25].

The Kinect sensor can be used to evaluate gait and postural changes in these patients to facilitate powerful physical rehabilitation for the gait impaired stroke patients. The concept of exergames also adopts machine learning and game AI to support the integration of natural human motion and exercise for the elderly. This patient demographic may require continuous physiotherapy but may not necessarily have prescriptions for these in the community unless they are coming home from the hospital setting. Exergames can be set-up for the home environment coupled with tele-mentoring platforms whereas serious games can facilitate the remote monitoring via AI-supported platforms with alerting systems to enable remote mentoring/monitoring and feedback.

A limitation of these games is that they are restricted to the analysis of only impairment-

level-assessments that focus on non-occluding movements, and are unable to diagnose deficits of gross and fine movement in the hand and foot [3]. The games are not usually suitable for all ages, however machine learning algorithms could be trained and programmed to ensure patient-age-targeted games are developed.

The Kinect system as a forebearer to the current Microsoft HoloLens 1 and 2 augmented reality platform also enables registration and joint detection algorithms for the recognition of the shoulder, neck and arm joints. This allows for a multidimensional analysis of shapes, sizes, postures, and other motions to be characterized. Such a high dimensionality of datasets benefits from principal component analysis to determine torso joint position and pose estimation together with Fourier transforms for temporal dynamic changes [3].

Others have also proposed the naïve Bayesian nearest neighbor algorithms such as the Eigen-joint to combine multiple information to aid classification [26]. Others also introduced novel approaches for human action recognition, termed the histogram of 3D joint locations, which extracts spherical coordinates of the position of joints as histograms [27].

Depth image extraction of the 3D position of joints has been achieved using Gaussian kernels to compute confidence scores. Kinect skeleton joint features have also benefitted from algorithms such as support vector machines, K-nearest neighbors, NITE algorithms, etc. [3, 28] These algorithms have been used to model diseases where movement is impaired and to facilitate exercises for physical rehabilitation including motion such as helping patients with performing jumping jacks, arm circles, arm curls, motion of limbs and hips in ways that minimize strain on their joints [3].

Others have also employed kinematic filtering, point cloud segmentation, cylinder model fitting, hierarchical coordinate systems, Z-score normalization, Kalman filters, fuzzy logic, skeleton normalization, and dynamic time warping as well as trajectory recognition to characterize diseased motion for machine learning [3, 29–33]. Furthermore, deep neural networks were also leveraged by Han and colleagues to identify recovery states

and postural correction [33]. Other methods such as zero-velocity crossing to locate starting and ending points for human motion segmentation have also been discussed in the literature [31, 34].

The ability to judge depth by an algorithm is a challenging area that has taken decades of research to develop [1, 3, 17]. Local Ternary Directional Pattern feature descriptors have enabled encoding of depth maps [35]. These feature descriptors can be fed into support vector machine classifiers. This highly performant ML approach outperforms existing descriptors such as HOG, PHOG and CENTRIST, reducing the non-linear feature dependency of classification on the depth map by a support vector machine [3]. Others have proposed HOG-based depth motion maps computing the difference between two frames captured by the Kinect sensor and aggregating the difference over time [36]. Diseases where depth data have proven useful for physio rehabilitation include multiple sclerosis, stroke, wrist and hand injury, and cerebral palsy [37–39]. Machine learning algorithms that have been implemented include K-means, neural networks, fuzzy logic, PCA, decision forest and Histogram of Oriented 4D Normal among others [3, 31, 36, 38, 39].

Estimating pose from the exact body contours of an individual has also been achieved by Xia et al. [27] These and many other areas will continue to evolve over the coming decades to provide robust methods of supporting physical rehabilitation. However, what must also be addressed is the emotional and psychological aspects of physical rehabilitation that can limit patients' compliance, and this is an area in which machine learning feedback systems must evolve to support.

In a gamification setting, there is a need to also provide psychological support when the patient does not successfully complete a task to the desired level or reach their goal. Encouragement is imperative and is one of the crucial roles that a human physiotherapist can provide, together with empathy and motivation. The response to a patient's perceived failure must be measured, but also objectively analyzed so that an appropriate level of support can be adjusted to limit

psychological stress. Machine learning algorithms using natural language processing like the Amazon Alexa, Kinect and others could facilitate this encouragement dynamic to support patients in settings when in-person physiotherapists are not readily available. Physiotherapists cannot be available 24/7 and patient access to services may also be limited due to lack of funding or during situations such as the SARS-COV-2 pandemic, which require social distancing measures.

For exercise assessments, Capecci and colleagues used Histogram Semi-Markov models (HSMM) to monitor and evaluate the conduct of the rehabilitation exercises comparing it to Dynamic Time Warping and reporting superior performance in favor of HSMM to correlate better with clinical scores [31].

Companies in this space are summarized in Table 1 and include SWORD Health, a Portuguese-based start-up that has designed an evidence-backed solution using wireless trackers to obtain biofeedback from the patients' movements [40, 41]. An AI-based physical therapist guides the patient through a few exercises. In an insurance-powered global healthcare system, the physiotherapist together with a remote team of human clinicians could be based in any country and could be incentivized to check in periodically with the patient. This tele-platform approach ensures 24 h support for the patient taking into consideration the different time-zones for both patients and physiotherapists [40, 41].

Other companies such as Physera have an online application and eHealth facilities aiming to facilitate face-to-face physiotherapy between patients and therapists [5]. Physitrack is another company that uses video-platform based methods for orthopedics and neurology patients to enable them to consume several hours of video and learn from various exercises [5]. VeraHealth provides a platform that enables community physiotherapy services to be achieved [5].

Another is the OpenPose platform and library, which is an open-source digital library for musculoskeletal physiotherapy [5, 6]. The OpenPose system enables a computer-aided rehabilitation platform created in C++ for posture detection. It facilitates an online platform for physio-based

Table 1 A summary of some of the companies that are involved in AI for physiotherapy

Apps/ Companies	Purpose	Machine learning technologies behind it.	Clinical examples	Future	Evidenced based medicine	Reference
Orbbec	Fall detection. Patient monitoring. In-home rehab. Disabled assistance.	Depth Vision 3D Camera	Provides data for highly effective analysis of the rehabilitation process to allow for proper treatment, as well as determine the best course of action.	Same technique used in advanced robotics. Likely that robotics will be an integral part of future physiotherapy.	√	[17]
Sword Health	Home-based physical rehabilitation	Wireless motion trackers attached to the body. “digital therapist” gives feedback in real time; “Home-based Rehabilitation with a NovelDigital Biofeedback System versus Conventional In-person Rehabilitation after Total Knee Replacement: a feasibility study” <i>Nature: Scientific Reports</i> , July 26, 2018	Clinically-validated programs work for all the major MSK issues, at any point in the journey: prevention, acute conditions, chronic pain, and post-surgical recovery.		√	[40, 41]
PhysiApp by Physitrack	Telemedicine. Tracks patient exercise performance, adherence, and outcomes	It also features anatomical imagery with 3D content provided by Primal Pictures.	“Physitrack is scientifically proven to increase home exercise adherence and patient confidence”	Now 80,000+ healthcare professionals and 1 million patients per year in more than 100 countries. Future growth...	no	[5, 59]
Kaia Health	Motion Coach. Front-facing camera of phone. An on-screen wireframe model guides patients through exercises, and a proprietary machine-	“AI-movement coach.” Computer vision. Patients can use a chat interface to consult with a physical therapist about the exercises.	Back pain. Pain management. <i>Kyle Wiggers, ‘Kaia Health’s app uses AI to alleviate back pain,’ VentureBeat, September 20, 2018.</i>	See footnote below	√	[60]

(continued)

Table 1 (continued)

Apps/ Companies	Purpose	Machine learning technologies behind it.	Clinical examples	Future	Evidenced based medicine	Reference
	learning algorithm evaluates patient feedback and adapts the exercise program to the patient's needs					
BetterHelp	CBT/psycho-therapy for many psychological conditions.	Chat function with therapists. Face to face. No AI involved.	BetterHelp offers access to licensed, trained, experienced, and accredited psychologists (PhD / PsyD), marriage and family therapists (LMFT), clinical social workers (LCSW/ LMSW), and board licensed professional counsellors (LPC)	Future fusion of CBT, coaching, and physiotherapy AI apps. The betterHelp app would likely benefit from an initial AI assortment or feedback or monitoring feature with embedded AI.	✓	[61]
Bionik Laboratories, InMotion	InMotion	Michal Prywata, "Bois Are Becoming Highly Skilled Assistants in Physical Therapy," <i>VentureBeat</i> , October 15, 2017.	Use of AI and Telerobotics for physiotherapy from severe impairment to high level strengthening and coordination.	These mechanical physical therapists can work collaboratively with humans: The robots help the patient fine-tune each movement, while the therapists help the patient translate these improvements into greater function.	✓	[62]
PHIO by EQL	"Phio access - clinically-led digital MSK triage support tool. Phio Engage - tailored training programs	Responsive (user interface) UI replicates human interaction	AI in digital physiotherapy triage and clinical decision support, with a specific focus on decision trees (DT) and probabilistic modelling (PM).	unclear	A web published systematic review of other evidence was presented on the companies website https://www.eqi.ai/clinical-excellence	

Well Health	AI-embedded mobile app.	<p>Self-Management of Chronic Neck and Back Pain using artificial neural networks that were trained on a total of 300 sets of training samples. Once the initial weighting of the MLP-ANN was trained by the samples provided by the experts, the back-propagation algorithm. Model validation was performed by comparing the AI-generated exercise program with the expert-generated exercise program.</p>	<p><i>Chronic musculoskeletal neck and back pain, self-perceived benefits, n = 161. This study demonstrated the positive self-perceived beneficiary effect of using the AI-embedded mobile app to provide a personalized therapeutic exercise program</i></p>	√	[63]
KINECT by Microsoft (discontinued)	Discontinued by Microsoft	<p>KINECT sensor associated with upper limb joint position and Artificial Neural Network classification algorithms an evaluation and correction of the patients' exercises can be done, helping therapists to improve the effectiveness of training sessions. To improve the patient motivation a set of Kinect serious games for upper limbs training assures 3D data capture of the patients' joints and data storage in a remote database</p>	<p>Microsoft Research Cambridge-12 Kinect (MSRC-12) dataset</p>	Discontinued	Discontinued [18]

(continued)

Table 1 (continued)

Apps/ Companies	Purpose	Machine learning technologies behind it.	Clinical examples	Future	Evidenced based medicine	Reference
Internet of Things, generally on fitness wearables	A survey for smart fitness that focuses which covers IoT-based solutions for the physiotherapeutic domain and looks at the impacts of artificial intelligence and social- IoT			An EBM attempt was made with a survey performed.		[64]

training of patients. It integrates with computer vision frameworks such as OpenCV, OpenCL, and Caffe for image-based training and live image rendering, where a recorded video can be analyzed and feedback can be provided to the patient [5].

AI in Physiotherapy Education and Use of Simulation for Educating Physiotherapists

The role of physiotherapists has continued to evolve over many decades and provides specially trained and regulated subspecialists in most medical institutions. Physiotherapists often work as part of a multidisciplinary team in various hospital settings ranging from medicine to surgery, outpatient clinics as well as the community health centers and clinics, general practitioners' surgeries, and sports and exercise clubs. There is an unmet need for the education of physiotherapists and their continuous professional development to cater for any subsequent knowledge gaps [4, 42].

Correspondingly, in-situ simulation settings enable senior physiotherapy supervisors to train junior therapists in a safe, controlled environment that mimics the actual clinical situation of interest. This equips them with the necessary practical skills to manage often very complex patients needing a multi-system and multi-disciplinary approach to physiotherapy, occupational therapy, speech and language therapy, to support areas such as breathing training as well as mobilization.

Many algorithms have been developed that are finding their way into high fidelity simulators to heighten the learning experience of physiotherapists. Algorithms that facilitate the recognition and classification of human activities in video sequences, which segment out background distractions, etc. can be used for human action recognition to train physiotherapists to solve a particular problem [5].

Convolutional neural networks, recurrent neural networks and other deep neural networks, and computer vision together with support vector machines are also commonly used to analyze pose and motion recognition. Moreover, random

forests have also been employed for action recognition and can be incorporated to break down specific tasks to enable the learner to make improvements, analyze complications, and improve patient safety [3, 4, 10, 11].

The delivery of physiotherapy education can be achieved remotely. The objective of one study was to assess the effect of simulated learning environments (SLE) on educators' self-efficacy in student supervision skills [43]. Significant improvements in clinical supervision and self-efficacy change scores were seen in SLE participants compared to control participants in three domains of self-efficacy: (1) talking to students about supervision and learning styles ($p = 0.01$); (2) adapting teaching styles for students' individual needs ($p = 0.02$); and (3) identifying strategies for future practice while supervising students ($p = 0.02$) [43]. A facilitated debrief, where most of the learning occurs, can also be augmented by artificial intelligence. For an in-depth treatment of AI in medical education please see the relevant chapter in this compendium of works.

Another intriguing area of physiotherapy education is in advanced gaming immersion utilizing augmented reality and virtual reality settings. In one study, the authors mention that advanced AR/VR and gaming technologies can authenticate simulated learning experiences that are close to real-life conditions, problems, and applications, something that could revolutionize training [42]. The GAMEPHARM platform was described where researchers targeted and consolidated the findings from three distinct pillars to produce a prototype and blueprint for physiotherapists looking to design systems and platforms for rehabilitation education using games. These pillars included: (a) the requirements of the application area in question; (b) the current state-of-the-art and emerging directions in game-based professional education and training; and (c) existing applications of game-based learning in the field of healthcare [42]. Their team proposed an idea and design of augmented reality modules embedded alongside the game simulation and virtual world environment. This is aimed at providing participants with an authentic immersive simulated game scenario. As ML algorithms continue

to evolve, an additional pillar would be to investigate the progression of game AI to support educational physiotherapy-based rehabilitation.

AI and Physiotherapy Education

Physiotherapy education also involves the troubleshooting of problems presented by other less experienced members of the team by a more experienced team leader and educator who provides support for the process of physiotherapeutic diagnostics [44]. The approach is to streamline the number of physical tests that the trainee physiotherapist will perform to reach a clinical diagnosis, as multiple tests can be considerably uncomfortable for the patient and an inefficient use of the therapist's time. In one study, the researchers compiled the anatomy of the neuromuscular system and developed an AI-based tool (called PhysIt), which creates an interactive visualization and diagnostic system to assist the trainee physiotherapist [45]. The efficacy of their diagnostic and visualization tool was successfully evaluated and significantly decreased the number of imprecise candidate diagnoses of patients with physiotherapeutic needs [44, 45].

AI for Robotic Assisted Physiotherapy

Another area with potential for growth is AI-based robotic physiotherapy to support patient rehabilitation. Over recent decades there has been a keen interest in robotic-assisted physiotherapy to improve patient mobility. Either the robots behave in an assistive way or they completely take away the added joint stress that the patient undergoing early rehabilitation is being placed under. According to some physiotherapists who oppose this approach, one disadvantage may be that having the robot take away the added joint stress means that the patient fails to learn how to use that joint correctly, which can inevitably delay their recovery. However, in one study, the team retrospectively identified 107 cases of new cerebral stroke who were allocated into 2 groups. A robotic physiotherapy group where 36 patients

underwent 30 sessions of robotic physiotherapy twice a week using the Lokomat, compared with 71 patients undergoing a 5 times per week conventional physiotherapy program [46]. The Modified Ashworth Spasticity Scale (MASS), Brunnstrom Recovery Scale (BRS), Functional Independence Measure (FIM), and Functional Ambulation Categories (FAC), as well as the Berg Balance Scale (BBS) were used as evaluation parameters. Cognitive evaluators included the Mini Mental State Examination and Short Form Health surveys. The percentage of parameters at discharge relative to pre-treatment values demonstrated superior improvements in FIM, MMSE, in the RT group ($P < 0.05$), suggesting that robotics could be incorporated into current approaches for physical rehabilitation to aid cognitive augmentation of human-guided physiotherapy [46]. The authors did not report on the combination of both human and robotic methods.

Another aspect of robotic physiotherapy is the layer of education that is not available in traditional physiotherapy curricular. Twenty-first-century healthcare will continue to incorporate machine learning methods and the physiotherapy-educator is uniquely positioned to have these skills in their armamentarium to unlock approaches that decades ago would have been considered infeasible. It may be that the future physiotherapist will program to delegate all the repetitive reasoning tasks to an AI agent, such as a robot, to facilitate interpersonal patient-physiotherapy educational development. They may therefore need to educate themselves and learn about how the AI powered robot will interact with the patient and ways to augment its behavior to provide the right support for the patient.

AI for Physio-assisted Activity of Daily Living Monitoring

In one publication, the ATHENA smart process management system aims to bring in advanced planning to improve daily activity and cognitive impairment in patients requiring complex physiotherapy such as for stroke and dementia conditions [2]. The platform leverages AI-based

planning and scheduling to design timed sequences of activities that solve a problem in each environment. Their web-based framework also analyzes and facilitates ambient assisted living for decision support and for the daily care of patients, which enables the design of patient-specific personalized activities of daily living. Technical staff can be organized and coordinated for patients needing a multi-disciplinary input from the entire medical team such as nurses, physicians and physiotherapists as well as the presence of the infrastructure required for their rehabilitation, such as monitors, the need for laboratory support, etc.

AI and Virtual Reality for Physiotherapy and Rehabilitation

Recently others have developed off-the-shelf video game and virtual reality platforms for rehabilitation [47–49]. One such work describes an avenue for integrating virtual reality and AI-based game tracking methods to improve the effectiveness of hand rehabilitation. Prototyped using the LEAP motion sensor as the input device, the tracking cameras leverage the Oculus virtual reality headset for interactive gaming developed on the Unity Game Engine [47]. Their preliminary studies reported good efficiencies in implementation for human subjects in terms of hand physical therapy [48]. Others have also described a similar approach for leveraging virtual reality for upper limb rehabilitation in patients with neurological disease and another termed VirtualPT for home care and elderly patient rehabilitation [49].

AI for Physiotherapy-assisted Sensory and Balance Training

Falls are a common problem affecting about 35% of all community-based individuals aged 65 or older with increasing risk the older the individual becomes [1, 3, 9–11]. Physiotherapists also help patients with their balance and one such platform to train patients with vestibular pathologies to improve their ability to sense their surroundings

and reduce their risk of falling has been discussed by Condron and colleagues [50]. They developed a dual task programmable platform called the Chattecx Balance System (CBS), a computerized force platform that is aimed at facilitating both balance performance and cognition. They looked at twenty healthy elderly patients with a mild increase in fall risk. Artificial intelligence through wearables can be used to support patients who are losing their balance [50].

In a randomized controlled trial the effectiveness of Ai Chi was compared with conventional water-based exercise to improve balance [9]. The researchers used balance performance measures in areas such as the excursion and movement velocity and balance control assessments such as the limit of stability test as well as the Berg Balance Scale(BBS). They reported a significant increase in the anteroposterior direction of motion after receiving Ai Chi ($p = 0.005$ for excursion, $p = 0.013$ for velocity) but not in conventional water-based exercise. The Ai Chi group demonstrated significantly better results than the control group ($p = 0.025$).

AI for Assisted Wheelchair Users and Assisted Mobility Support

A significant proportion of the global population are wheelchair users and in the developed world they receive significant support and advocacy to make their lives easier in terms of mobility. The need for a wheelchair may arise due to a variety of neurological and trauma-related conditions such as spinal cord injury and stroke leading to quadriplegia or paraplegia, multiple sclerosis, and cerebral palsy. These individuals with disability could benefit from AI. Over recent decades, artificial intelligence has been considered for allowing physiotherapy management and mobility support for wheelchair users [1, 7, 10]. Novel engineering approaches have enabled the development of the smart wheelchair, which has evolved since George Westinghouse's 1914 invention of the manually controlled electric wheelchair into a power-driven automated innovation that enables physiotherapy treatment compatibility [7]. Current

innovations include autonomous AI-based advanced speech/voice recognition powered by the Android operating platform that have been incorporated into the design of the smart wheelchair [7, 10]. This now allows the patient to move freely in a highly secured environment using embedded flex sensors to create an awareness of the wheelchair's surroundings. It has the capability to navigate and reduce any strain on an individual and the design can facilitate the needed physical therapy.

Other assistive mobility devices include alternative and autonomous systems as well as augmentative mobility training devices, those that make use of human machine interfaces and brain computer interfaces and electromyography signal processing as well as those for biped station positioning(the ability to support the accomplishment of certain tasks on two legs) [10].

AI for Inattention and Hemi-neglect Training

Inattention monitoring and rehabilitation after a stroke requires advanced physiotherapy and constant patient prompting. This enables an individual who has hemi-neglect and hence restricted use of one side of their body to learn to either slowly improve, recover or gain the ability to compensate using the other side [12]. Inattention mechanisms are poorly understood, but has been studied in motor vehicle drivers and usually include distraction and fatigue but in some patients such as those with stroke the issues are organic in nature leading to hemi-neglect. [51] Currently Constraint-Induced Movement Therapy has shown some promise, but in a systematic review none of the tools designed for managing neglect had evaluated physical performance [12]. One possible area where AI could augment and supplement inattention retraining involves the constant prompting required for the patient to re-learn how to use the other side. A combination of natural language processing, robotics and computer vision can facilitate supplementary prompting with tools such as Amazon Alexa and video cameras helping with the monitoring of the patient's state

especially in scenarios where physiotherapists may not be available on a 24/7 basis.

Despite the growing recent interest in the utility of robotics for autonomous logistical support to aid physiotherapy and rehabilitation, we are yet to have widespread use of fully dedicated humanoid robots to support patient physiotherapy [52]. Inattention retraining could be an area that benefits from the robotic space where robots can be trained to help prompt and guide patients who are not using one side of their bodies.

Another area where AI can benefit inattention retraining involves the use of wearables and smart apps to induce stimuli that could prompt the patient to use the side that they are neglecting. Given the shortage of physiotherapists, as described by several different countries due to the economic landscape, out-of-hours physiotherapy support for conditions such as stroke patients with inattention would perhaps benefit from emerging AI-based technological support.

AI for Respiratory Physiotherapy Management

Clinical disease support systems can improve patient outcomes and decision-making processes. Current approaches are leveraging Internet of Things devices that can collect and stream a vast amount of physiological data from ventilators and other medical devices [53]. Machine learning models offer the unique potential and capability to model these datasets and derive insights into how ventilator-dependent patients can benefit from improved clinical processes and outcomes [15, 16, 53]. Physiotherapists specializing in respiratory and ventilator support can benefit from automated and advanced data analytics for decision support to provide them with insights on managing acute respiratory distress syndromes as illustrated in the recent SARS-COV-2 pandemic. Occasionally neurological diseases that affect the respiratory system like Guillain-Barré syndrome can also need respiratory physiotherapy support.

In one study authors looked at the respiration rate (RR), a vital observation and marker of respiratory disease, and the most widely adopted

techniques used to monitor it. Their work highlighted the many drawbacks clinicians' face when analyzing this marker. They used non-invasive contactless infrared thermography to predict respiratory rate [54]. Their system utilized a thermal camera to monitor variations in nasal temperature during continuous respiration, tracking regions of interest like the nostrils and head motion. Machine learning and computer vision algorithms made use of "Histogram of oriented gradients" and "Support vector machines" (SVM) to build a Breath Detection Algorithm (BDA). Algorithmic performance of this BDA was validated on 150 breathing signals through its precision, sensitivity, spurious cycle rate, and missed cycle rate value. Reported accuracies were 98.6%, 97.2%, 1.4%, and 2.8% respectively. The parameters obtained from the BDA were then passed into the k-Nearest Neighbor (k-NN) and SVM classifiers to determine whether the human test participants had abnormal or normal respiration, which they classified as bradypnoea or tachypnoea. The patterns were visualized using t-stochastic neighbor embedding (t-SNE). Reported validation accuracies were stated as 96.25% and 99.5% with training accuracies 97.75% and 99.4% for SVM and k-NN classifiers, respectively [54].

AI for Community Physiotherapy and Care

In many parts of the world, with a specific example being Bangladesh where 415 physiotherapists have been deployed over a 35-year period to support a 150 million population, practices are concentrated in the cities with limited access of services for the rural communities [55]. A 2005 World Bank report revealed that a significant number of up to 77% of the impoverished can only access local pharmacists or medicine sellers for their ailments due to poor accessibility to healthcare facilities and high consultation fees [55]. The shortage of community physiotherapists means innovative approaches combined with culturally viable physiotherapy services could improve the quality of life of individuals with

physical and emotional disabilities needing rehabilitation.

Machine learning, robotics and most of the previously discussed technologies could be employed to support community physiotherapy services in countries where there is an economic shortage of physiotherapists to supplement the deficits in services.

AI for Cognitive Impaired Patients Needing Physiotherapy and Rehabilitation

In many patients, especially those with neurological diseases such stroke, Parkinson's disease, multiple sclerosis, Alzheimer's disease, traumatic brain injury, etc., there is a clear need to support and manage cognitive deficits. However, these cognitive deficits also affect the patient's ability to be compliant with the physiotherapy rehabilitation regimens that could improve their quality of life and functional capacity.

Authors have discussed the use of symbiotic neuroprosthetics and myoelectric control as well as the use of brain machine interface technologies for perioperative medicine to augment cognition [56]. The use of an artificially cognitive machine developed for cognitive rehabilitative exercise was based on the machine's indications of when cognitive function was impaired [56]. The challenge though is that cognitive dysfunction is very dynamic and requires constant reassessment. Such algorithms must be trained to dynamically monitor the patient's state and alter their response to adjust the rehabilitation regimen or know when to involve or alert a physiotherapist to improve the patient's rehabilitation needs.

AI for Functional and Feedback Systems in Physiotherapy

In one study aiming to help with home-based physiotherapy, the identification of the patient's pose was assessed using the PoseNet. [6] In their study, the system used seven exercises with two sides for more variety, identifying posture and

pose, and generated a scoring function on the analyzed input image. The main idea of their desktop application system using a 3D Avatar is based on the ability to detect the body's key points using pose estimation techniques by extracting key points from different exercises and building a database file that stores the information. This was then compared with the patient's key points and a minimum accuracy of 80% was reported [6].

AI in Smart Watches and Wearables for Physiotherapy

Physiotherapy is recommended in patients experiencing arthritic shoulder pathology, as conservative management approaches are often initially recommended over immediate surgery [3, 5, 6, 8]. However post-surgical rehabilitation also requires physiotherapeutic regimens implemented over a period of weeks to months to ensure that the patient regains optimal function. Current limitations mean that there are no objective measures for unsupervised home physiotherapeutic regimens and exercise programs that allow adherence monitoring. In one proof of concept study aiming to resolve this disparity, a commercial smartwatch was utilized for a home physiotherapy study [8]. Using an evidence-based rotator cuff protocol, data from a six-axis inertial sensor were gathered from the active extremity. A framework called the activity recognition chain formed the base upon which four supervised learning algorithms, K-nearest neighbors, Random Forest, Support Vector Machine Classifiers, and Convolutional Neural Networks, were trained and optimized for classification of the exercises. Algorithmic performance was evaluated using a fivefold by-subject and temporally stratified cross-validation with a categorical classification accuracy of 94%. The convolutional neural network achieved the greatest performance in accuracy reported as 99.4%, but a lower accuracy score of 88.9% was observed when the algorithm was applied to an algorithmically naïve/unseen dataset [8]. Others like McGirr and collaborators have reported similar accuracies using 3-axis gyroscopes on shoulder exercise data in

classification problems and have reported accuracies of 86–91% [57]. Another study by Pan and colleagues used multiple synchronized inertial sensors mounted to the chest, upper arm and one wrist for a classification accuracy of 96.9 on five different shoulder exercises [58].

Future of AI in Physiotherapy

Physiotherapy and rehabilitation are an expensive service needing considerable financial resources that appear well suited to a private insurance-based healthcare system. However, the evolution of AI hopes to reduce costs and support physiotherapists with emerging technologies that will help improve the service delivery and make physiotherapy widely assessable to poorer communities. Technologies like robotic physiotherapy and tele-platform-based physiotherapy, with the right financial incentive and government backing, could support global service delivery of therapies for a large population of individuals. Other challenges that need to be overcome will be the mentality and fear of AI taking over jobs. As pragmatists one can appreciate that the technology has potential, however at such an early stage of development, an AI's role must seek to augment and supplement rather than completely replace the patient-clinician/physiotherapist relationship that has been established over several centuries. Although, as realists we should welcome the use of AI in areas where repetitiveness diminishes the efficiency of a therapist including record keeping, which can be automated using AI. But we must then seek avenues to educate in areas where physiotherapists can also gain the essential skills necessary to be at the forefront of this wave of innovation for better patient outcomes. It is here that we see potential for going forward in the future. Research areas that have evolved include Kinect-based gamification approaches for physiotherapy and the use of augmented reality for patient monitoring. Areas in physiotherapy education also show promise in terms of knowledge-based service delivery and provision of diagnostics as well as feedback for patients receiving physiotherapy across the

community. Wearables, smart sensors, and natural language processing platforms have also gained prominence in patient rehabilitation and will continue to evolve. An area where we see a confluence of technologies between application of natural language processing systems such as the Amazon Alexa with computer vision and robotics is in the cognitive aspects of physiotherapeutic rehabilitation, training, and feedback for patients with pathological deficits such as stroke, gait abnormalities and movement disorders. More robust machine learning algorithms will continue to evolve to help manage the more complex cognitively impaired patients that require continuous physiotherapeutic support.

References

1. Wei W. Using sensors and ai to enable on-demand virtual physical therapist and balance evaluation at home. University of San Diego Thesis. 2020.
2. Hidalgo E, Castillo L, Madrid RI, García-Pérez Ó, Cabello MR, Fdez-Olivares J. ATHENA: smart process management for daily activity planning for cognitive impairment. In: Bravo J, Hervás R, Villarreal V, editors. Ambient assisted living. IWAAL 2011 Lecture Notes in Computer Science, vol. 6693. Berlin/Heidelberg: Springer; 2011.
3. Rashid F, Suriani N, Nazari A. Kinect-based physiotherapy and assessment: a comprehensive review. Indones J Electr Eng Comput Sci. 2018;11(3):1176–87.
4. Ramanandi V. Role and scope of artificial intelligence in physiotherapy a scientific review of literature. Int J Adv Sci Res. 2021;6(1):11–4.
5. Godse S, et al. Musculoskeletal physiotherapy using artificial intelligence and machine learning. Int J Innov Sci Res Technol. 2019;4:11.
6. Hassan H, et al. Automatic feedback for physiotherapy exercises based on PoseNet, vol. 2(2). Informatics Bulletin, Helwan University; 2020.
7. Rani E, Niranjana R. Novel engineering of smart electronic wheelchair with physiotherapy treatment compatibility. In: Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC); 2019. p. 1–4. <https://doi.org/10.1109/I-SMAC47947.2019.9032603>.
8. Burns DM, Leung N, Hardisty M, Whyne CM, Henry P, McLachlin S. Shoulder physiotherapy exercise recognition: machine learning the inertial signals from a smartwatch. Physiol Meas. 2018;39(7):075007.
9. Ku PH, Chen SF, Yang YR, Lai TC, Wang RY. The effects of Ai Chi for balance in individuals with chronic stroke: a randomized controlled trial. Sci Rep. 2020;10(1):1201.
10. Martins MM, Santos CP, Frizera-Neto A, Ceres R. Assistive mobility devices focusing on Smart Walkers: classification and review. Robot Auton Syst. 2012;60(4):548–62.
11. Carmeli E. Physical therapy for neurological conditions in geriatric populations. Front Public Health. 2017;5:333.
12. Hellström BVK. Treatment and assessment of neglect after stroke – from a physiotherapy perspective: a systematic review. Adv Physiother. 2008;10(4):178–87.
13. Teikari P, Pietrusz, A. Precision strength training: data-driven artificial intelligence approach to strength and conditioning. SportRxiv May 20. <https://doi.org/10.31236/osf.io/w734a>. 2021.
14. Achttien R, Staal JB, van der Voort S, et al. Exercise-based cardiac rehabilitation in patients with coronary heart disease: a practice guideline. Neth Hear J. 2013;21(10):429–38. <https://doi.org/10.1007/s12471-013-0467-y>.
15. Vitacca M, Barbano L, Vanoglio F, Luisa A, Bernocchi P, Giordano A, Panerini M. Does 6-month home caregiver-supervised physiotherapy improve post-critical care outcomes? Am J Phys Med Rehabil. 2016;95(8):571–9.
16. Gosselink R. Physiotherapy in respiratory disease. Breathe. 2006;3(1):30–9.
17. Calin A, Coroiu A. Interchangeability of Kinect and Orbbec sensors for gesture recognition. In: IEEE 14th International Conference on Intelligent Computer Communication and Processing (ICCP); 2018. p. 309–15. <https://doi.org/10.1109/ICCP.2018.8516586>.
18. Deboeverie F, Roegiers S, Allebosch G, Veelaert P, Philips W. Human gesture classification by brute-force machine learning for exergaming in physiotherapy. In: IEEE Conference on Computational Intelligence and Games (CIG); 2016. p. 1–7.
19. Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) 2005, vol. 1; 2005. p. 886–93.
20. Mohan A, Papageorgiou C, Poggio T. Example-based object detection in images by components. IEEE Trans Pattern Anal Mach Intell. 2001;23(4):349–61.
21. Ke Y, Sukthankar R. PCA-SIFT: a more distinctive representation for local image descriptors. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004 CVPR 2004, vol. 2; 2004. p. II-506–13.
22. Belongie S, Malik J, Puzicha J. Shape matching and object recognition using shape contexts. IEEE Trans Pattern Anal Mach Intell. 2002;24(4):509–22.
23. Zhu Q, Yeh M, Cheng K, Avidan S. Fast human detection using a cascade of histograms of oriented gradients. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06) 2006, vol. 2; 2006. p. 1491–8.
24. Zhang W, Zelinsky G, Samaras D. Real-time accurate object detection using multiple resolutions. In: IEEE 11th International Conference on Computer Vision 2007; 2007. p. 1–8.

25. Lowe D. Distinctive image features from scale-invariant keypoints. *Int J Comput Vis.* 2014;60(2): 91–110.
26. Yang X, Tian Y. EigenJoints-based action recognition using Naïve-Bayes-Nearest-Neighbor. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops 2012; 2012. p. 14–9.
27. Xia L, Chen C, Aggarwal J. View invariant human action recognition using histograms of 3D joints. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops 2012; 2012. p. 20–7.
28. De Rosario H, Belda-Lois J, Fos F, Medina E, Poveda-Puente R, Kroll M. Correction of joint angles from kinect for balance exercising and assessment. *J Appl Biomech.* 2014;30(2):294–9.
29. Staab R. Recognizing specific errors in human physical exercise performance with Microsoft Kinect. Master's Theses Proj. Reports. California Polytechnic State University, San Luis Obispo. 2014.
30. Li S, Pathirana P, Caelli T. Multi-kinect skeleton fusion for physical rehabilitation monitoring. In: 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society; 2014. p. 5060–3.
31. Capecci M. Physical rehabilitation exercises assessment based on Hidden Semi-Markov Model by Kinect v2. In: IEEE-EMBS International Conference on Bio-medical and Health Informatics (BHI) 2016; 2016. p. 256–9.
32. Lee J, Hsieh C, Lin T. A Kinect-based Tai Chi exercises evaluation system for physical rehabilitation. In: IEEE International Conference on Consumer Electronics (ICCE) 2014; 2014. p. 177–8.
33. Han S, Kim H, Choi H. Rehabilitation posture correction using deep neural network. In: 2017 IEEE International Conference on Big Data and Smart Computing (BigComp); 2017. p. 400–2.
34. Lin JF-S, Kulic D. Automatic human motion segmentation and identification using feature guided hmm for physical rehabilitation exercises. Workshop on Robotics for Neurology and Rehabilitation, IEEE International Conference on Intelligent Robots and Systems 33-6 2011. 2011.
35. Shen Y, Hao Z, Wang P, Ma S, Liu W. A novel human detection approach based on depth map via Kinect. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops 2013; 2013. p. 535–41.
36. Nahavandi D, Hossny M. Skeleton-free task-specific rapid upper limb ergonomic assessment using depth imaging sensors. Proc IEEE Sensors. IEEE, Piscataway, N.J. 2016;1–3. <https://doi.org/10.1109/ICSENS.2016.7808687>
37. Collins J, Warren J, Ma M, Proffitt R, Skubic M. Stroke patient daily activity observation system. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2017. p. 844–8.
38. Bakar M, Samad R, Pebrianti D, Mustafa M, Abdullah N. Computer vision-based hand deviation exercise for rehabilitation. In: IEEE International Conference on Control System, Computing and Engineering (ICCSCE); 2015. p. 389–94.
39. Sosa G, Sánchez J, Franco H. Improved front-view tracking of human skeleton from Kinect data for rehabilitation support in Multiple Sclerosis. In: 20th Symposium on Signal Processing, Images and Computer Vision (STSIVA); 2015. p. 1–7.
40. Correia F, Nogueira A, Magalhães I, Guimarães J, Moreira M, Barradas I, Molinos M, Teixeira L, Tulha J, Seabra R, Lains J, Bento V. Medium-term outcomes of digital versus conventional home-based rehabilitation after total knee arthroplasty: prospective, parallel-group feasibility study. *JMIR Rehabil Assist Technol.* 2019;6(1):e13111.
41. Correia F, Nogueira A, Magalhães I, et al. Home-based rehabilitation with a novel digital biofeedback system versus conventional in-person rehabilitation after total knee replacement: a feasibility study. *Sci Rep.* 2018;8(11299):2018.
42. Pappa D, Papadopoulos H. A use case of the application of advanced gaming and immersion Technologies for Professional Training: the GAMEPHARM training environment for physiotherapists. *Electr J e-Learning.* 2019;17(2):157–70.
43. Holdsworth C, Skinner E, Delany C. Using simulation pedagogy to teach clinical education skills: a randomized trial. *Physiother Theory Pract.* 2016;32(4):284–95.
44. Rowe M. Artificial intelligence in clinical practice: implications for physiotherapy education. *Open Physio J.* 2018.
45. Mirsky R, Hibah S, Hadad M, Gorenstein A, Kalech M. “PhysIt” – a diagnosis and troubleshooting tool for physiotherapists in training. *Diagnostics.* 2020;10:72.
46. Dundar U, Toktas H, Solak O, Ulasli A, Eroglu S. A comparative study of conventional physiotherapy versus robotic training combined with physiotherapy in patients with stroke. *Top Stroke Rehabil.* 2014;21(6): 453–61.
47. Pillai M, Yang Y, Ditmars C, Subhash H. Artificial intelligence-based interactive virtual reality-assisted gaming system for hand rehabilitation. In: Proc SPIE 11318, Medical Imaging 2020: imaging informatics for healthcare, research, and applications, 113180J; 2020.
48. Shahmoradi L, Almasi S, Ghobbi N, Gholamzadeh M. Learning promotion of physiotherapy in neurological diseases: design and application of a virtual reality-based game. *J Educ Health Promot.* 2020;9:234. Published 2020 Sept 28
49. Heiyanthuduwa TA, Amarapala KWNU, Gunathilaka KDVB, Ravindu KS, Wickramarathne J, Kasthurirathna D, editors. VirtualPT: virtual reality based home care physiotherapy rehabilitation for elderly. 2020 2nd International Conference on Advancements in Computing (ICAC), 10–11 Dec 2020; 2020.
50. Condon JE, Hill KD, Physio GD. Reliability and validity of a dual-task force platform assessment of balance performance: effect of age, balance

- impairment, and cognitive task. *J Am Geriatr Soc.* 2002;50:157–62.
51. Dong Y, Hu Z, Uchimura K, Murayama N. Driver inattention monitoring system for intelligent vehicles: a review. *IEEE Trans Intell Transp Syst.* 2011;12(2):596–614.
52. Lonner J, Zangrilli J, Saini S. Emerging robotic technologies and innovations for hospital process improvement. In: Lonner J, editor. *Robotics in knee and hip arthroplasty*. Cham: Springer; 2019.
53. Rehm GB, Woo SH, Chen XL, Kuhn BT, Cortes-Puch I, Anderson NR, et al. Leveraging IoTs and machine learning for patient diagnosis and ventilation management in the intensive care unit. *IEEE Pervasive Comput.* 2020;19(3):68–78.
54. Jagadev P, Giri LI. Human respiration monitoring using infrared thermography and artificial intelligence. *Biomed Phys Eng Express.* 2020;6(3):035007.
55. Ellangovin M. Innovations in community physiotherapy Field Actions Science Reports [Online], Vol 2 | 2009, Online since 17 September 2010, connection on 06 June 2021. 2009.
56. Anderson D. Artificial Intelligence and Applications in PM&R. *Am J Phys Med Rehabil.* 2019;98(11):e128–9.
57. McGirr K, Harring S, Kennedy T, Pedersen M, Hirata R, Thorborg K, Bandholm T, Rathleff MS. An elastic exercise band mounted with a Bandcizer can differentiate between commonly prescribed home exercises for the shoulder. *Int J Sports Phys Ther.* 2015;10:332–40.
58. Pan J, Chung H, Huang J. Intelligent shoulder joint home-based self-rehabilitation monitoring system. *Int J Smart Home.* 2013;7:395–404.
59. Bennell K, Marshall CJ, Dobson F, Kasza JB, Lonsdale C, Hinman RS. Does a web-based exercise programming system improve home exercise adherence for people with musculoskeletal conditions? *Am J Phys Med Rehabil.* 2019;98(10):850–8.
60. Toelle TR, Utpadil-Fischler DA, Haas KK, et al. App-based multidisciplinary back pain treatment versus combined physiotherapy plus online education: a randomized controlled trial. *npj Digit Med.* 2019;2:34.
61. Marcelle E, Nolting L, Hinshaw SP, Aguilera A. Effectiveness of a multimodal digital psychotherapy platform for adult depression: a naturalistic feasibility study. *JMIR Mhealth Uhealth.* 2019;7(1):e10948.
62. Prywata M. Bots are becoming highly skilled assistants in physical therapy. *VentureBeat.* 2017;15. <https://venturebeat.com/2017/10/15/bots-are-becoming-highly-skilled-assistants-in-physicaltherapy/>
63. Lo W, Lei D, Li L, Huang D, Tong K. The perceived benefits of an artificial intelligence-embedded mobile app implementing evidence-based guidelines for the self-management of chronic neck and back pain: observational study. *JMIR Mhealth Uhealth.* 2018;6(11):e198. <https://doi.org/10.2196/mhealth.8127>.
64. Farrokhi A, Farahbakhsh R, Rezazadeh J, Minerva R. Application of Internet of Things and artificial intelligence for smart fitness: a survey. *Comput Netw.* 2021;189:107859.



Parastu Rahgozar

Contents

Introduction	1810
Definition of Rehabilitation	1810
Benefits of Rehabilitation	1810
Phases of Rehabilitation	1810
Areas of Rehabilitation Medicine	1811
Physical Therapy	1811
Speech Therapy	1811
Neuropsychology	1811
Occupational Rehabilitation	1812
General Applications of AI in Rehabilitation	1812
Robotics in Rehabilitation	1812
AI in Assessment and Decision Support Systems	1814
AI in Rehabilitation Prognosis	1814
AI Wearable Monitoring Devices	1815
AI in Virtual Reality and Serious Games	1815
Summary	1816
References	1816

Abstract

Technological advancements in the past decade, especially in the field of Artificial Intelligence (AI), have influenced almost every industry, and the field of medicine is not an exception. From robots taking care of time-consuming, repetitive tasks in hospitals to

rapid cancer diagnosis methodologies developing every day, it is visible that AI has potential to help further the medical discipline.

AI in rehabilitation has broad usability, such as assisting in the rehabilitation session, evaluating the treatment progress (decision support), and providing prognosis regarding risk of complications or success of the treatment.

In this chapter, firstly rehabilitation and its specialties will be explained followed by a thorough explanation of why AI can be helpful in rehabilitation. Furthermore, different applications of AI in this field will be discussed. The

P. Rahgozar (✉)
KTH Royal Institute of Technology, Stockholm, Sweden
e-mail: parastu@kth.se

chapter also brings some examples from recent studies and state-of-the-art research.

Keywords

Rehabilitation · Machine learning · Physical therapy · Artificial intelligence · Robotics

Introduction

Advancements in technology, especially in the field of AI and robotics, have influenced almost every industry. Looking at a broad field like medicine, the impacts of AI is visible in various shapes and forms. From robots taking care of time-consuming, repetitive tasks in hospitals to rapid cancer diagnosis methodologies developing every day [1], it is visible that AI has potential to help further the medical discipline. In some fields this effect is already visible, while in others it is yet to be proven.

Rehabilitation is an inseparable part of most medical procedures and treatment plans. Therefore, it is vital to look for solutions to automate the tedious processes in the field and instead utilize the regained power on tasks that require the most attention. The need of a powerful processing and analyzing tool results in reaching out for machine learning algorithms in rehabilitation as well. ML algorithms are great tools to be used with large amounts of multiparameter data. This data can be used to personalize care services for individuals and help medical staff in decision making [2].

There are a variety of subjects in AI that can be integrated with rehabilitation. Diagnosis and accurate classification of types of disorders with similar symptoms, unifying medical records in medical databases, decision support systems for therapists, assisting them in rehabilitation session, and prediction of treatment outcome are part of the accomplishments in this field. Examples of such applications are further mentioned in this chapter.

Definition of Rehabilitation

The concept of rehabilitation is generally referred to as a series of tasks and actions which aims for a better physical and social life for the patient. According to the WHO, rehabilitation is defined

as: “a set of interventions designed to optimize functioning and reduce disability in individuals with health conditions in interaction with their environment” [3].

The aim of rehabilitation is to help a patient with a disability or injury to overcome the pressure and be able to adapt to a new life routine as much as possible. In the ideal way, the patient will be able get back to their normal life routines. This goal includes all different aspects of an individual’s life [4].

Benefits of Rehabilitation

Rehabilitation can help with decreasing the burden of a wide range of health issues. This includes chronic or acute diseases, injuries, traumas, or other illnesses. It can also have a tremendous role when it comes to controlling symptoms or severeness of a disease or it can be widely used as a step after surgeries so that patients gain back muscle strength or prevent complications after surgery. A vast number of patients do not only need physical support but also need to consult with a psychologist and receive advice on how to bear with certain inevitable life situations, such as accepting that the patient might not get back pre accident levels of performance.

One of the specialties in rehabilitation in medicine is physical medicine. This specialty focuses on quality-of-life enhancement of patients with a health-related disability and how to provide the maximum level of social integrity for the patient.

Phases of Rehabilitation

Despite a general misunderstanding, rehabilitation is not limited to one single phase in the care procedure. Usually, the plan and execution of a rehabilitation process is spread into separate periods [4]:

1. Acute period of illness or injury
2. Early rehabilitation (intensive care in clinic or hospital)
3. Treatment and rehabilitation phase, in collaboration with the patient

4. Shifting focus more towards rehabilitation and decreasing direct care
5. Focus on vocational rehabilitation with prevention of secondary complications
6. Patient's state support with a long-term mindset

Generally, the efforts in rehabilitation are put forward toward two major goals. The process is done either toward improving a certain deficiency or reducing the effects of a certain deficiency.

According to statistics [4], one-third of patients do not require any rehabilitation after they face an injury or illness, and they are able to go back to their normal life routines after receiving clinical treatment. However, the rest of the patients will need assistance to gain full recovery and be able to adapt again to their life routine, both physically and socially.

Areas of Rehabilitation Medicine

When it comes to rehabilitation, often only the physical regards come to mind, although a full concept includes several different aspects to consider. There are several ways to define rehabilitation, and therefore, there can be different views on how to categorize areas in rehabilitation medicine [4].

In order to reach the ideal outcome in the rehabilitation process, there are several subsections that rehabilitation is divided into. These subsections are physical therapy, speech therapy, neuropsychology, and occupational therapy.

Although in an actual medical case it is not feasible to separate physical medicine into subgroups, but for an ideal and optimal outcome for a patient, all four aspects must be taken into account.

For instance, a patient with an injury from an accident needs different kinds of support. They require physical therapy to help them attain the physical function they may have lost because of the injury. They may also require psychological support as well to recover from the mental distress.

Physical Therapy

The World Confederation for Physical Therapy (WCPT) has defined physical therapy as: "providing services to people and populations to develop,

maintain and restore maximum movement and functional ability throughout the life-span. Physiotherapy includes the provisions of services in circumstances where movement and function are threatened by the process of ageing or that of injury to disease" [5].

Physical therapy is divided into two different general approaches, diagnostic and therapeutic. The diagnostic approach is used to distinguish different sorts of disorders based on their specific symptoms while a therapeutic view strives to improve the physical symptoms of the disorders which affects the patient's life [6].

Speech Therapy

A speech disorder is defined as any sort of problem a person can have with fluency, voice, or sound while they speak. There are many different disorders that can be categorized as a speech disorder or have speech difficulties as a symptom of other illnesses [7].

Speech and language therapy consists of a series of methods and approaches such as verbal, behavioral, and natural language baselines. These therapy sessions may also include behavioral therapy and response training. In this area, the effort is made to help the person to use their speech capabilities more productively and try to improve the ways of communication. Moreover, therapists in this field work on strengthening the patient's ability to produce the correct sounds in speech and make the conversation more understandable and less tiresome for the speaker [7].

Neuropsychology

Neuropsychological disorders are characterized as issues with increased deficits in behavioral, cognitive, or psychological matters. These issues are mostly caused by a trauma or injury occurred to the brain [8]. Part of neuropsychological illnesses can also occur due to natural aging, such as Alzheimer's or dementia [9]. The science of neuropsychology aims for helping the patients to improve their psychological state and retrieve their social skills to blend in better in society [10].

One of the main challenges that is part of neuropsychologists' work is the uniqueness of most cases. No two brains function the same, and no two personalities are identical. In addition, the therapy session is timely and often has to be done physically in a clinical environment [9].

Therefore, it can be quite demanding to tailor, personalize, and follow the goals and treatment planning for each individual patient [11]. In section "[AI in Rehabilitation Prognosis](#)" some examples of AI aids in this field will be discussed.

Occupational Rehabilitation

Occupational injuries and disorders are generally any abnormal condition that has happened related to a person's employment or because of certain exposures linked to work environment. An extensive number of occupational disorders are in musculoskeletal injuries category, and that is the reason why the initial rehabilitation programs were designed for this group. Currently, this area has broadened to different perspectives in health conditions, such as mental and neurological disorders [12].

The British Society of Rehabilitation Medicine (BSRM) describes occupational rehabilitation as "A process whereby those disadvantaged by illness or disability can be enabled to access, return to, or remain in, employment, or other useful occupation" [13].

General Applications of AI in Rehabilitation

After giving a good perspective of rehabilitation and its specialties, it is important to explain why and in which form artificial intelligence comes into the world of rehabilitation. In general, AI algorithms can be used in forms of supervised learning (learning with labels), semi-supervised learning (a mix of data with and without labels), unsupervised learning (learning without labels), and reinforcement learning (reward-based learning). Apart from the logic behind AI models, data is the most important component in AI. In fact,

data is the main resource used to build the models and make assumptions about them.

The methods and models defined in ML can be utilized in several areas of rehabilitation, such as classification, clustering, and prediction. For instance, recent research studies have worked on the application of machine learning algorithms in classification and recognition of human movements. In [14] Li et al. use support vector machines (SVMs) and K-nearest neighbors to classify hand postures in different exercises related to rehabilitation. Related to the same field, there have been cooperative studies done on the power of different classifiers when it comes to differentiating between functional upper limb movements and walking patterns [15]. Moreover, there are models used for prediction of patients' prognosis which can be a great asset in treatment planning as well. SVMs and K-nearest neighbors have also shown promising results in prediction of patients' recovery. This method has been compared to the common practices currently used in the field which is Activities of Daily Living Clinical Assessment Protocol (ADLCAP) to identify patients with potential for improvement and earlier home discharge. The proposed method has shown a better performance than ADLCAP [16].

But the applications of AI do not end there. Recent advancements have shown that AI algorithms can not only learn actions and tasks performed by therapists but they can also then adjust the difficulty level of the tasks to suit every individual patient's needs [17]. As an example, Shirzad et al. [18] have demonstrated an assessment between K-nearest neighbors, neural networks, and discriminant analysis methods. In this study the mentioned algorithms have been compared based on their ability to adjust the difficulty level of tasks to the patients' motor performance and physiological features accordingly.

Robotics in Rehabilitation

Previously, the main application of robots in rehabilitation was limited to assistive purposes, meaning that robots were involved in the process to

ease some daily routines for the patient. The first rehabilitation robot was designed in the 1980s and it was defined as “Any automatically operated machine that is designed to improve movement in persons with impaired physical functioning” [19, 20].

In the past four decades, the means of using a robot in rehabilitation tasks have improved significantly. Tasks in physical rehabilitation can be repetitive for a therapist but a robot is not prone to fatigue and can handle intense, repetitive tasks for a longer period of time than a human. This factor by itself makes robots a great resource in occupational rehabilitation [17].

The main source of input for these human-made systems is the data that they receive through sensors. The decision taken based on this data can be both fully automated and operated by another person [21]. The capabilities of robots are very much determined by the level of complexity of the learning algorithms used and also the number of sensors used to capture data from the environment [17]. For instance, a simple sensor-based system to capture room temperature can be elevated to a complex system by adding humidity sensors or air pressure sensors. In the new setting, there is more available data to process, and thus more distinct predictions can be made based on more detailed data.

In the world of rehabilitation, robots can have an exceptional effect when it comes to assistive help to care givers, or to the patients both in physical and cognitive aspects.

According to [21], robotic applications can be found in different tasks nowadays. For instance, robots can help a patient gain strength in their muscles after an injury or assist them to have smoother movements.

Learning from Demonstration

Learning from Demonstration (LfD) is a set of machine learning techniques that makes a robot able to learn certain tasks by imitating an expert [22]. Then, the robot would be able to reproduce the learned tasks based on the patterns and rules learned from the demonstrator. The demonstrator in this case would be the therapist and the AI model would be the imitator. This technique is

the most useful in settings such as therapy clinics and patients’ houses where there can be lack of enough resources such as computers and therapists [23]. One of the main benefits of LfD techniques is that, thanks to the advanced AI and sensors, the system can assess the patients’ physical performance and tailor the tasks suitable for their readiness level. LfD systems can be both semi-autonomous and fully autonomous, where the therapist can be present as an overseer to the performance of the robot and interrupt if errors are made. This feedback can then be used by the robot to learn from.

One area that LfD has been applied in is cases such as cerebral palsy and stroke. These conditions can have symptoms such as loss of motor function and reduced mobility. According to [24] in patients with Cerebral Palsy (CP), at least one CP symptom can cause issues in cognitive and linguistic development especially in children under the age of 18 and affect their life expectancy. The study suggests a semiautonomous assistance robot which first uses Gaussian mixture model (GMM) to copy the performed trajectory, and later uses Gaussian mixture regression (GMR) to produce a statistically similar process. Meanwhile, patient’s motion will be controlled with a PID (proportional-integral-derivative) controller.

Gait Rehabilitation

One of the areas in which robotics has been a big aid to therapists is gait rehabilitation. Usually, the methodology in this subject involves the application of exoskeletons for paralyzed patients in need of gait rehabilitation.

While the traditional rehabilitation methods are still common, there are various studies reporting considerable improvements in results of gait rehabilitation done by robots [25]. This is a good example to show the effectiveness of assistive robots, but it is important to emphasize that there are also practical differences between the generations of the mentioned robots.

The initial generation of robotic exoskeletons in this field uses predetermined, fixed patterns to train gaits, which means the patient do not have an active role in the process and therefore the process of recovery will take much more time.

Although the recent studies suggest a multimodal human-robot interaction system (HRI) [26] which offers interactions with the patient both cognitively (cHRI) and physically (pHRI), these methods are applied by adapting the treatment process to each patient's current state.

There are also rehabilitation robots that use artificial intelligence as a model to personalize treatment for maximum outcome [27]. One study suggests using a training robot that uses two exoskeletons for the lower limb area, in combination with an adaptive admittance model. The model is adaptive which means that its law is defined based on a sigmoid function and reinforcement learning algorithms. In this approach both passive and active training have been experimented where the active training focuses on personalizing the movements and the passive part verifies the control strategy in the model [27].

AI in Assessment and Decision Support Systems

As it was mentioned in the introduction, assessment of a patient's state plays a big role in determining further steps of the treatment plan. The current practice includes therapists making an assessment based on their experience and mostly empirical knowledge and intuition. They must infer diagnosis based on symptoms and patient's self-reporting which does not include any explicit numbers or quantitative data. The factors taken into account are mostly descriptive and unmeasurable. Also, there is great risk that therapists are not always available to give an assessment periodically. The mentioned issues are why AI can play an important role in medical assessment since it can be accessible without limitations, offers a measurable, numerical system, and therefore provides an intelligent decision support system based on a more precise assessment [28].

Rapid developments in the field of data processing are some of the reasons behind huge success of artificial intelligence in different knowledge areas. In case of medical information systems, a model requires sufficient, reliable sources to accurately organize and represent

patient data. There are also different approaches when it comes to decision-making models [29].

AI in Rehabilitation Prognosis

Since a great number of treatments planning in rehabilitation depend on the outcome of the previous treatment stage, one of the factors that can play an important role in helping therapists is the accurate prediction and analysis of so far applied treatment or prognosis or chronic diseases. Also, such outcome can assist therapists to foresee the patient's assistive needs.

There are an extensive number of prediction models in ML that can be implemented either with supervised or unsupervised learning. Some of the most common ones in nonmedical studies are regression models, decision trees, and random forests (prediction models) [30].

In [30] the subject of memory dysfunction of patients with Alzheimer's disease, the performance of ML algorithms regarding prediction of cognitive health status classification has been examined. In this study, KNNs, decision trees, Logistic Regressions (LR), Multilayer Perceptrons (MLP), Naive Bayes (NB), two varieties of Random Forest (RF100 and RF500), Radial Basis Functional Networks (RBF), and Support Vector Machines (SVM) were utilized for a binary classification of yes/no answers in a health status questionnaire. Results demonstrated an overall better performance for LR in this case.

Unlike Alzheimer, some illnesses have symptoms in early childhood such as CP. Although there is no permanent cure for CP, studies on the treatment of this neurological disorder have never stopped. Nonsurgical approaches such as visual therapy have shown encouraging improvements in CP patients which works on specific visual dysfunctions. In the previous studies, AI has been an aid to detecting and diagnosing different diseases. But when it comes to evaluation of visual improvement in CP children, there have not been much computational advancements.

The suggested methodology in [31] uses ML techniques in both diagnostic and prognostic aspects. By capturing images of the CP patients'

eyes and taking advantage of image processing methods, they have been able to not only classify the eye images into four different eye abnormalities but also to extract quantified information from the images for prognosis evaluation. The calculated percentage of improvement is also used to analyze the effectiveness of the visual therapy in general.

AI Wearable Monitoring Devices

Since AI is built upon data, the availability of resources to provide data for AI models is one of the most important tasks in any field where AI is involved. Wearable devices and intelligent systems such as mobile phones are some of the best examples when it comes to data sources. Almost every person owns a smart device which is supplied with several sensors and processors to gather data. One benefit is that smart watches are available in different price ranges and tech specifications for purchase. Also, smart watches or wristbands are carried by the patient for an extended amount of time and the data collection does not require any specific interaction between the patient and the device. Meaning that patient does not need to input data manually. This continuous process of collecting data is a great source of information especially health-related, physiologic, and behavioral material. In general, these factors can help therapists get a better understanding of ongoing rehabilitation therapy. Wearable devices can be used by elderly people, patients with a history of stroke, or patients under supervision for an extended period of time. Athletes and many other groups of people can also gain from this.

Fall Detection

One of the common uses of wearable devices is their build-in fall detection system which has a simple but important purpose of notifying care takers in case the patient loses balance or falls for any reason [32]. The logic in some of these systems is designed in a way to distinguish between other types of movements and an actual fall with the aid of machine learning algorithms. These algorithms learn the behavioral patterns

while they do normal daily tasks and use the information to lower the number of false positive alarms in this case.

Monitoring Purposes

Another example is the practicality of wearable monitoring devices for patients with cardiovascular problems. Studies [2] have proven that cardiac rehabilitation programs are the second highest effective prevention method that has shown promising results when it comes to treatment approaches. The program requires participants to complete the monitoring and data collection over a period of time which often is demanding and results in participants dropping out or not completing the program successfully. The idea of changing the programs setting to a home-based approach and collecting data via wearable devices has been a proper solution to this matter. As it was mentioned before, the biggest benefit of using wearable devices is the constant data collection which gives the caretaker a thorough view of the patient's status and makes the choice of AI advantageous over questionnaires and reports filled by patients.

AI in Virtual Reality and Serious Games

There are many types of exercise in physical therapy which are advised to be done after an injury, and they do not require a supervision of a therapist at all times. However, it is important for a smooth recovery and lower risk of complications that patients follow the instructions and executes the tasks correctly. Therefore, application of virtual reality and gamification was introduced in the world of rehabilitation. One of the benefits of this technology is the ease of usage and the fact it does not need supervision of a therapist at all times.

The setup can vary in different available products. Some can have a display with instructions, images, or a simulated environment, while there are products that use VR headsets for visual purposes. Based on its purpose some systems have a controller stick that users can hold, and the sensors can capture motion and movements. In [33] such a system is used for hand therapy, one of the most common types of rehabilitation before or

after a surgery. The suggested setup in [33] consists of capturing data from the sensors on a headband and the controller in the user's hand. The data will then be applied in a machine learning model to improve the hand's gestures and analyze the effectiveness of the therapy. The final purpose of this project is not only to help the patient complete their tasks but also to provide some entertainment with games.

The most interesting part of this study is how they have used both passive and real time classification in the system [33]. When the patient completes one task, an SVM algorithm will choose the next suitable exercise based on their performance score. Additionally, the patient is provided with a feedback message where it is explained how accurate their grip was compared to the desired outcome. A KNN algorithm has been implemented to take care of this process [33]. The outcome of this research can be applied for other rehabilitation purposes or to design more exercises using the same setup.

Summary

In this chapter, the definition of rehabilitation medicine, rehabilitation medicine's necessity in medical practice, and the different aspects of rehabilitation through AI have been discussed. There are numerous benefits of using AI in other industries which have helped them become more autonomous and at the same time apply their manpower in a smarter way. The same approach has been applied in rehabilitation with AI and it has had great achievements according to numerous studies in the field. The state-of-the-art algorithms and considerable amounts of data are two important factors behind the success of AI. A variety of examples about the influence of AI in rehabilitation are also mentioned in this chapter.

References

- Huang S, et al. Artificial intelligence in cancer diagnosis and prognosis: opportunities and challenges. *Cancer Lett.* 2020;471:61–71.
- De Cannière H, Corradi F, Smeets CJ, Schouteten M, Varon C, Van Hoof C, Van Huffel S, Groenendaal W, Vandervoort P. Wearable monitoring and interpretable machine learning can objectively track progression in patients during cardiac rehabilitation. *Sensors.* 2020;20(12):3601.
- World Health Organization. <https://www.who.int/news-room/fact-sheets/detail/rehabilitation>. Accessed 30 Jan 2021.
- Kolar P. Clinical rehabilitation. Alena Kobesová; 2014.
- Kumar SP. Physical therapy: past, present and future-a paradigm shift. *J Phys Ther.* 2010;1(2):58–67.
- Binkley J, et al. Diagnostic classification of patients with low back pain: report on a survey of physical therapy experts. *Phys Ther.* 1993;73(3):138–50.
- Danubianu M, Pentiu SG, Schipor OA, Nestor M, Ungureanu I. Distributed intelligent system for personalized therapy of speech disorders. In: 2008 the third international multi-conference on computing in the global information technology (ICCGI 2008) 2008 Jul 27. IEEE. p. 166–70.
- Wilson BA. Neuropsychological rehabilitation. *Annu Rev Clin Psychol.* 2008;4:141–62.
- Sirsat YA, i Badia SB, Fermé E. AI-rehab: a framework for ai driven neurorehabilitation training the profiling challenge. *InHealthinf.* 2020;845–853.
- Robinson G, Weekes G. Rehabilitation in clinical neuropsychology. 2007. <https://doi.org/10.4324/9780203783924>.
- Wilms I. Using artificial intelligence to control and adapt level of difficulty in computer-based, cognitive therapy—an explorative study. *J Cyberther Rehabil.* 2011;4:387.
- Gross DP, Haws C, Park J. Occupational rehabilitation. In: Gellman MD, editor. *Encyclopedia of behavioral medicine.* Cham: Springer; 2020. https://doi.org/10.1007/978-3-030-39903-0_101938.
- Ross J. Occupational therapy and vocational rehabilitation. Wiley; 2013.
- Leightley D, Darby J, Li BH, McPhee JS, Yap MH. Human activity recognition for physical rehabilitation. *IEEE Syst Man Cybern.* 2013;261–6. <https://doi.org/10.1109/Smc.2013.51>.
- McLeod A, Bochniewicz EM, Lum PS, Holley RJ, Emmer G, Dromerick AW. Using wearable sensors and machine learning models to separate functional upper extremity use from walking-associated arm movements. *Arch Phys Med Rehabil.* 2016;97(2):224–31. <https://doi.org/10.1016/j.apmr.2015.08.435>.
- Zhu M, Zhang Z, Hirdes JP, Stolee P. Using machine learning algorithms to guide rehabilitation planning for home care clients. *BMC Med Inform Decis Mak.* 2007;7(1):1–3.
- Fong J, Ocampo R, Gross DP, Tavakoli M. Intelligent robotics incorporating machine learning algorithms for improving functional capacity evaluation and occupational rehabilitation. *J Occup Rehabil.* 2020;30:362–70.
- Shirzad N, Van der Loos HFM. Adaptation of task difficulty in rehabilitation exercises based on the user's motor performance and physiological responses. In: 2013 IEEE 13th international conference on rehabilitation robotics (ICORR). 2013.

19. Zhao Y, Liang C, Gu Z, Zheng Y, Wu Q. A new design scheme for intelligent upper limb rehabilitation training robot. *Int J Environ Res Public Health.* 2020;17(8):2948.
20. Reinkensmeyer DJ. Rehabilitation Robot. <https://www.britannica.com/technology/rehabilitation-robot>. Accessed 24 Jan 2021.
21. Luxton DD, Riek LD. Artificial intelligence and robotics in rehabilitation. In: Brenner LA, Reid-Arndt SA, Elliott TR, Frank RG, Caplan B, (eds), *Handbook of rehabilitation psychology*. American Psychological Association. 2019;507–20. <https://doi.org/10.1037/0000129-031>.
22. Atkeson CG, Schaal S. Robot learning from demonstration. In: Proceedings of the fourteenth international conference on machine learning (ICML '97), vol. 97. Morgan Kaufmann; 1997. p. 12–20.
23. Najafi M, Adams K, Tavakoli M. Robotic learning from demonstration of therapist's time-varying assistance to a patient in trajectory-following tasks. In: 2017 international conference on rehabilitation robotics (ICORR). IEEE; 2017. p. 888–94.
24. Klein T, Gelderblom GJ, Witte LD, Vanstipelen S. Evaluation of short-term effects of the IROMEC robotic toy for children with developmental disabilities. *IEEE Int Conf Rehabil Robot.* 2011;2011:1–5.
25. Krebs HI, Hogan N, Aisen ML, Volpe BT. Robot-aided neurorehabilitation. *IEEE Trans Rehabil Eng.* 1998;6(1):75–87.
26. Gui K, Liu H, Zhang D. Toward multimodal human–robot interaction to enhance active participation of users in gait rehabilitation. *IEEE Trans Neural Syst Rehabil Eng.* 2017;25(11):2054–66.
27. Bingbing G, et al. Human–robot interactive control based on reinforcement learning for gait rehabilitation training robot. *Int J Adv Robot Syst.* 2019;16(2):1729881419839584.
28. Lee MH, Siewiorek DP, Smailagic A, Bernardino A. Opportunities of a machine learning-based decision support system for stroke rehabilitation assessment. *arXiv preprint arXiv:2002.12261.* 2020.
29. Ceccaroni L, Subirats L. Interoperable knowledge representation in clinical decision support systems for rehabilitation. *Int J Appl Comput Math.* 2012;11(2):303–16.
30. Bergeron MF, et al. Episodic-memory performance in machine learning modeling for predicting cognitive health status classification. *J Alzheimers Dis.* 2019;70:277–86.
31. Illavarason P, Renjit JA, Kumar PM. Medical diagnosis of cerebral palsy rehabilitation using eye images in machine learning techniques. *J Med Syst.* 2019;43(8):1–24.
32. Casilar E, Oviedo-Jiménez MA. Automatic fall detection system based on the combined use of a smartphone and a smartwatch. *PLoS One.* 2015;10(11):e0140929.
33. Pillai M, Yang Y, Ditmars C, Subhash H. Artificial intelligence-based interactive virtual reality-assisted gaming system for hand rehabilitation. In: *Medical imaging 2020: imaging informatics for healthcare, research, and applications*, vol. 11318. International Society for Optics and Photonics; 2020. p. 113180J.



João Gustavo Claudino, Daniel de Oliveira Capanema, and
Paulo Roberto Pereira Santiago

Contents

Introduction	1819
Advances	1820
Potential Trends	1820
Future Challenges	1822
References	1823

Abstract

The potential of AI for Sports Medicine has been highlighted in the literature. However, few results are found when the command line

(“artificial intelligence”) AND (“sports medicine”) was applied in three of the main databases. These findings show how much we still have to improve in this area of knowledge and this chapter is going to start this process by presenting advances, potential trends, and future challenges of AI in Sports Medicine.

J. G. Claudino (✉)

Research and Development Department, LOAD CONTROL, Contagem, Minas Gerais, Brazil

School of Physical Education and Sport – Laboratory of Biomechanics, Universidade de São Paulo, São Paulo, São Paulo, Brazil

e-mail: [jooao.gustavo@loadcontrolapp.com;
claudinojo@usp.br](mailto:jooao.gustavo@loadcontrolapp.com; claudinojo@usp.br)

D. d. O. Capanema

Computing Department, Federal Center for Technological Education of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

P. R. P. Santiago

School of Physical Education and Sport of Ribeirão Preto – LaBioCoM Biomechanics and Motor Control Laboratory, Universidade de São Paulo, Ribeirão Preto, São Paulo, Brazil

e-mail: paulosantiago@usp.br

Keywords

Analytics · Fitness · Illness · Injury · Performance · Recovery · Adaptation, Industry 4.0 · Prediction · Evolution

Introduction

One of the main outcomes of the 4th Industrial Revolution, also known as Industry 4.0 is the large amount of data produced. Industry 4.0 is characterized by “cyber-physical systems” – systems that integrate computing, networking, and physical processes, as well a multitude of

technologies (e.g., internet of things, smart sensors) that generated large datasets and leveraged the artificial intelligence (AI) algorithms [1–6].

Recent manuscripts have highlighted the potential of AI for the Sports Medicine area [7, 8]. On the other hand, when applying the command line (“artificial intelligence”) AND (“sports medicine”) in 3 of the main databases, 47 results are found (at January/2021: Pubmed = 14; Web of Science = 04; Scopus = 29). Among them, only two were original research manuscripts published in a peer-reviewed journal and inside of the Sports Medicine area [9, 10]. First, the University at Buffalo’s Orthopedics Department created an expert system to assist patients with self-diagnosis of knee problems and to thereby facilitate referral to the right orthopedic subspecialist with Bayesian versus Heuristic method for diagnostic decision support. However, neither approach truly outperformed the other, and accuracies were not statistically and significantly different [9]. Second, Deep Learning was used for detection of complete anterior cruciate ligament tear in patients among 18–40 years old and a high performance in detection of complete ACL tears with over 96% test set accuracy was found [10].

These findings show how much we still have to improve in this area of knowledge and this chapter is going to start this process by presenting advances, potential trends, and future challenges of AI in Sports Medicine.

Advances

One of the first manuscripts on AI in the Sports Medicine found in the scientific literature was on the controversies and effectiveness of a 12-lead electrocardiogram (ECG) to detect cardiac diagnoses such as hypertrophic cardiomyopathy in young athletes [11]. The authors reported that barriers would be overcome in the near future with the use of AI to assist the decision-making process. In about 5 years, an automated detection of heart defects in soccer and basketball athletes based on ECG and artificial neural network (ANN) produced acceptable results of up to 98%

accuracy, 98% sensitivity, and 99% specificity [12]. Beyond the Sports Medicine field, an analysis of more than 180 thousand people revealed that an AI-enabled ECG acquired during normal sinus rhythm permits, even using an inexpensive, noninvasive, widely available point-of-care test, the identification of individuals with a high likelihood of atrial fibrillation. This result could have important implications for atrial fibrillation screening and for the management of patients with unexplained stroke. The improvement of these AI algorithms will increasingly allow the application of mobile devices, wearables that are widely used in sports settings also for these and other purposes [13]. Therefore, the advances of AI applications in Sports Medicine related to the ECG can already be recognized in a short period of 7 years (from 2012 to 2019) [11, 13].

The speed of advances is largely due to digital transformation that is becoming increasingly common in modern life and sports medicine [14]. This digitization also allows the integration with other important professionals of physical fitness. For example, Fitness Coaches or Personal Trainers using their scientific knowledge and practical expertise add up to the AI [14]. This integrated system is able to do not only measured parameters but also provide advice for progress or preventive measurements and help to create individualized training plans, by taking into account daily changes in performance [14, 15]. Furthermore, an integrated framework of load monitoring by a combination of smartphone, wearables, and point-of-care testing provides feedback that allows individual responsive adjustments to activities of daily living of athletes aiming to enhance performance and/or reduce the risk for overuse, injury, and/or illness [16] (Fig. 1).

Potential Trends

In the near future, the assessment of injury risk will be in real time to assist in decision making regarding the formation of the match and consequently the use of substitutes [17]. The integration of data collection with tracking systems including real-time microsensor inputs using AI algorithms



Fig. 1 AI in Sports Medicine advances

will be likely a key factor in soccer [17], but still needing improvement in data veracity [18]. Additionally, examining technological evolutions expected until 2038, the possibilities can result in something very close to the ideal for the Sports Medicine approaches. Let's look at this futuristic timeline proposed in "Countdown to the Singularity" [19]:

- 2030 ⇔ "AI passes the Turing test, meaning it can match (and exceed) human intelligence in every area."
- 2038 ⇔ "Everyday life is now unrecognizable – incredibly good and hyper virtual reality and AI augment all parts of the world and every aspect of daily human life."

Another potential trend in this field is the humanizing AI. According to Irani [20], even with a huge advance in data processing and computing via AI, it would not easily compare with the true potential of human intelligence. Creativity is one of the greatest AI challenges facing the human mind. Because creativity always comes as a surprise to us, if it was not surprising, we would not need it. Machines are not capable of creativity. Human minds can generate counterfactuals, imaginative flights, and dreams. Using AI surprises are welcome in some fields as art and music; however, surprises in medical diagnosis or treatment are unwelcome [20]. Nonetheless, the point of greatest concern is yet another [20], if AI is going to make clinicians better at caring for

humans, the data sets being used must be representative of society and not biased by sex, race, ethnicity, socioeconomic status, age, ability, and geography [21]. This need for representation is not only a data science issue, but also a moral one. In the absence of equal representation, society has already seen inequitable criminal justice sentencing, unfair hiring practices, and loan-risk determination, to name a few injustices [22]. We will talk more about this in the next topic.

AI also represents a great potential for healthcare applications in areas such as imaging and diagnoses, risk analysis, lifestyle management and monitoring, and health information management [23, 24]. This trend will allow a holistic approach applied 24 h per day, 7 days per week [16]. Thus 24/7 approach is expected to be facilitated by the availability of smart spaces (i.e., smart homes, smart cities), including training complex – facilities, athlete's home, type of transport, and/or the hotel, where mobile, wearable devices, and the internet of things will capture practically everything what athlete does inside and off the field. Such smart spaces may include, for instance, the possibility for monitoring quantity and quality of food and sleep, psychological, physical, physiological, medical and social demands of training/competition, energy expenditure, changes in body mass, hydration, and recovery levels outcomes. In addition, an integrated analysis including all these outcomes will be possible thanks to the advances in cloud computing, internet, big data, and specially AI that in



Fig. 2 24/7 approach: a potential trend

near future it will be expected to be at full levels of efficiency (Fig. 2).

Future Challenges

AI-related ethics is one of the greatest challenges for the future. Racine et al. [24] highlighted three potentially problematic aspects of AI use in healthcare and it certainly included the sports medicine area: (i) dynamic information and consent, (ii) transparency and ownership, and (iii) privacy and discrimination. The authors propose that AI-related ethical challenges may represent an opportunity for growth in organizations. The ethical data monetization is already today simultaneously an opportunity and a concern. The world is preparing to deal with this scenario via some regulations; the General Data Protection Regulations (GDPR) is a set of compliance

requirements applied in Europe and the Health Insurance Portability and Accountability Act (HIPAA) sets the standard for sensitive patient data protection in the United States. Some questions still need to be answered [24]:

- Does consent to evolving knowledge about treatment or prognosis, or to potential consequences formatters such as life insurance, need to be revisited? (Dynamic information and consent)
- Another important question concerns the ownership of the data upon which algorithms are developed. If the data come from an individual patient/athlete or an identifiable pool of individuals/athletes, will the intelligence developed be solely owned by the person (e.g., the club, the clinician, or the device company) stewarding its development? (Transparency and ownership)

- How much health information should patients/athletes be encouraged to share and how informed can they ever be regarding its future uses, possibilities for which seem to be growing every day? (Privacy and discrimination)

Furthermore, media enthusiasm for AI may pressure professionals and institutions involved to explore AI-based technology and adopt AI-informed practices prematurely without proper ethical guidance or empirical evidence of the validity/reliability of this technology [24, 25].

Finally, based on the open science approach, the dataset sharing should be performed aiming at area evolution both in academia and industry [26–28]. This culture of data sharing has already been strongly recommended in several areas [26, 27] including healthcare [28]. The authors reported that open science approaches should be adopted by AI Research and Development Departments in the healthcare domain, because, the use of “black-box” systems or the introduction of systems that have not demonstrated clinical effectiveness may not be acceptable approaches for the safety-critical healthcare context [28].

Before finalizing, we would like to do a brief statement that it is important to keep in mind that currently AI has the potential to classify, cluster, associate, visualize, and even select attributes (important variables). On the other hand, we cannot expect AI, for example, with Machine Learning techniques to be able to make extremely accurate predictions of such a complex phenomenon as a sport injury. Just remember that weather forecasts rely on extremely powerful computers and yet there are many errors. We must gather much more information to break the limit that AI currently presents. Larger datasets will give better results, but not close to the 99% that everyone wants. Initiatives such as that of the University of Waikato (New Zealand) as WEKA (Waikato Environment for Knowledge Analysis; an open source machine learning software in JAVA) [29] as well as that of the OpenAI (an AI research and deployment company with the mission to ensure that artificial general intelligence benefits all of humanity) [30] are so important for the future because these initiatives popularize AI and allow

progress in the search for real prevention, which would be to prevent the events before its occurrence. Therefore, so far what we have are classifications, prognostic, and prediction, but still searching for the significant capacity to really prevent a sport injury or other applications in sports medicine.

In conclusion, AI already presented deep advances in the field (e.g., ECG applications). The potential trends are related to allow real-time decision making in sports medicine, humanizing AI and a holistic approach applied 24/7. AI-related ethics is one of the greatest challenges for the future, thus actions and plans are warranted to solve it from now.

References

1. Buguin J, et al. Disruptive technologies: advances that will transform life, business, and the global economy. San Francisco: McKinsey Global Institute; 2013.
2. Winkelhaus S, Grosse EH. Logistics 4.0: a systematic review towards a new logistics system. *Int J Prod Res.* 2019;58:18–43.
3. Claudino JG, Capanema DO, de Souza TV, Serrao JC, Machado Pereira AC, Nassis GP. Current approaches to the use of artificial intelligence for injury risk assessment and performance prediction in team sports: a systematic review. *Sports Med Open.* 2019;5(1):28.
4. Ristevski B, Chen M. Big Data analytics in medicine and healthcare. *J Integr Bioinform.* 2018;15:20170030.
5. Kalid N, Zaidan AA, Zaidan BB, Salman OH, Hashim M, Muzammil H. Based real time remote health monitoring systems: a review on patients prioritization and related “Big Data” using body sensors information and communication technology. *J Med Syst.* 2017;42:30.
6. Reimer AP, Madigan EA. Veracity in big data: how good is good enough. *Health Inform J.* 2018. <https://doi.org/10.1177/1460458217744369>.
7. Kakavas G, Malliaropoulos N, Pruna R, Maffulli N. Artificial intelligence: a tool for sports trauma prediction. *Injury.* 2020;51:63–5.
8. Parker W, Forster BB. Artificial intelligence in sports medicine radiology: what’s coming? *Br J Sports Med.* 2019;53:1201–2.
9. Elkin PL, Schlegel DR, Anderson M, Komm J, Ficheur G, Bisson L. Artificial Intelligence: Bayesian versus Heuristic method for diagnostic decision support. *Appl Clin Inform.* 2018;9:432–9.
10. Chang PD, Wong TT, Rasiej MJ. Deep Learning for detection of complete anterior cruciate ligament tear. *J Digit Imaging.* 2019;32:980–6.

11. Chang AC. Primary prevention of sudden cardiac death of the young athlete: the controversy about the screening electrocardiogram and its innovative artificial intelligence solution. *Pediatr Cardiol.* 2012;33:428–33.
12. Adetiba E, Iweanya VC, Popoola SI, Adetiba JN, Menon C. Automated detection of heart defects in athletes based on electrocardiography and artificial neural network. *Cogent Eng.* 2017;4(1):1411220.
13. Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, Carter RE, Yao X, Rabinstein AA, Erickson BJ, Kapa S, Friedman PA. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet.* 2019;394(10201):861–7.
14. Rigamonti L, Albrecht UV, Lutter C, Tempel M, Wolfarth B, Back DA, Working Group Digitalisation. Potentials of digitalization in sports medicine: a narrative review. *Curr Sports Med Rep.* 2020;19:157–63.
15. Huang CC, Liu HM, Huang CL. Intelligent scheduling of execution for customized physical fitness and healthcare system. *Technol Health Care.* 2015;24 (Suppl 1):S385–92.
16. Dükking P, Achtzehn S, Holmberg HC, Sperlich B. Integrated framework of load monitoring by a combination of smartphone applications, wearables and point-of-care testing provides feedback that allows individual responsive adjustments to activities of daily living. *Sensors (Basel).* 2018;18(5):1632.
17. Nassis GP, Massey A, Jacobsen P, et al. Elite football of 2030 will not be the same as that of 2020: preparing players, coaches, and support staff for the evolution. *Scand J Med Sci Sports.* 2020;30:962–4.
18. Claudino JG, Cardoso Filho CA, Boullosa D, Lima-Alves A, Carrion GR, GianonI RLdS, Guimarães RdS, Ventura FM, Araujo ALC, Del Rosso S, Afonso J, Serrão JC. The role of veracity on the load monitoring of professional soccer players: a systematic review in the face of the Big data era. *Applied Sciences.* 2021;11 (14):6479. <https://doi.org/10.3390/app11146479>.
19. Diamandis P. Countdown to the Singularity. Available at: <https://medium.com/abundance-insights/countdown-to-the-singularity-52862e2c>. Accessed 28 Jan 2021.
20. Israni ST, Vergheze A. Humanizing Artificial Intelligence. *JAMA.* 2019;321(1):29–30. <https://doi.org/10.1001/jama.2018.19398>.
21. Caplan A, Friesen P. Health disparities and clinical trial recruitment: is there a duty to tweet? *PLoS Biol.* 2017;15(3):e2002040.
22. O’Neil C. Weapons of math destruction. How Big Data increases inequality and threatens democracy. New York: Penguin Books; 2016.
23. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science.* 2015;349(6245):255–60.
24. Racine E, Boehlen W, Sample M. Healthcare uses of artificial intelligence: challenges and opportunities for growth. *Healthc Manage Forum.* 2019;32(5):272–5.
25. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25(1):44–56.
26. McKiernan EC, Bourne PE, Brown CT, et al. How open science helps researchers succeed. *elife.* 2016;5: e16800.
27. Watson M. When will ‘open science’ become simply ‘science’? *Genome Biol.* 2015;16(1):101.
28. Paton C, Kobayashi S. An Open Science approach to artificial intelligence in healthcare. *Yearb Med Inform.* 2019;28(1):47–51. <https://doi.org/10.1055/s-0039-1677898>.
29. Frank E, Hall MA, Witten IH. The WEKA Workbench. Online appendix for “Data mining: practical machine learning tools and techniques”. 4th ed. Morgan Kaufmann; 2016.
30. Discovering and enacting the path to safe artificial general intelligence. Available at: <https://openai.com/>. Accessed 28 Jan 2021.

Index

A

Abductive diagnosis, 188–189
Ab initio tertiary structure prediction
 AlphaFold, 664
 fragment assembly, 664
 protein structure prediction method, 663
Absorption, distribution, metabolism, and excretion (ADME) assays, 638
Accident Compensation Corporation (ACC), 222
Accuracy, 1569, 1571, 1573
Accurate blood pressure measurements, 1151–1152
Acetic acid chromoendoscopy (AAC), 957
Active diagnosis, 183
Activity scoring, 1061
ACT-R models, 24
Acute Coronary Syndrome (ACS), 217
Acute ischemic stroke (AIS) therapy, 1504
Acute kidney injury (AKI) prediction
 explainability of models, 573
 external validity of models, 572–573
 implementation challenges, 573
 input features, 563–572
 ML algorithms, 572
 model performance, 572
 prediction timepoint and target period, 563
Acute lung injury (ALI), 1472
Acute lymphoblastic leukemia (ALL), 1404, 1431
Acute myeloid leukaemia, 1428–1431
Acute promyelocytic leukaemia, 1431
Acute respiratory distress syndrome (ARDS), 762, 1472
Acute stroke, 1225
Adaptive gamma correction method, 944
Adaptiveness, 413
Adaptive radiology interpretation and education system (ARIES), 333
Adaptive therapy, 345
Addiction, 1620, 1625
Adenine guanine cytosine and thymine (AGCT), 1392
Adenocarcinoma (LUAD), 763
Adenoma detection rate (ADR), 647, 654, 931, 968
ADME/T properties, 1061
Adoptive immune system, 1388

Advanced life support (ALS), 1472
Advanced Research Projects Agency Network (ARPANET), 210
Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER), 630
Adversarial autoencoder, 668
Adverse drug reaction (ADR), 1126–1128, 1135, 1136, 1138
Adverse Event Report System (FAERS), 219
Aerial Bone Mineral Density (BMD), 786
Africa, AIM
 Afya Pap, 620
 applications, 615–616
 CAD4TB, 620
 continental network model, 618
 costs and funding, 617
 data availability and quality, 616–617
 data ecosystem, 617–618
 development and capacity building, 618
 digital health foundation, 616
 digital health infrastructures, 617
 governance, regulations and ethics, 617
 governance and ethical approaches, 618
 history, 615
 infrastructure gaps, 618
 KEMSA, 620
 opportunities for AIM impact, 619
Afya Pap, 620
Ageing biomarker development, 1064–1065
Agent-based models, 1381
Age-related eye disease study (AREDS), 1529
Age-related macular degeneration (AMD), 211, 221, 1529, 1561–1564
 fundus photographs, 1520–1524
 generative adversarial network, 1543
 retinal fundus photographs, 1529
Agglomerative hierarchical clustering, 1131–1132
AI applications, 76, 1780
AI-assisted search, 259
AI-based approaches, 1390
AI-based decision support system (AI-DSS), 678

- AI based models, 582
 FFNN, 583–585
 LR, 583
 RBF, 586
 RNN, 585, 586
- AI-based software as a medical device (SaMD), 1544
- AiCure, 656
- AI-enabled Population Health, 619
- AI in medicine
 breaking boundary condition, 107
 correlations, 106
 gate keeper, 110
 goals, 111
 HyperTrak scanner, 103
 improvements, 109
 levels of expertise, 105
 limitations, 111
 machine learning, 108
 multi-sensor data, 104
 neural nets, 102
 prediction mode, 105
 risk and benefit analysis, 113
 self-evaluating mode, 106
 sociopath, 111
 tensors, 102
 terminology, 101
 training mode, 104
 transparency, 112
 validation, 109
- AI in Sports Medicine, 1820
 black-box systems, 1823
 Countdown to the Singularity, 1821
 healthcare applications, 1821
 12-lead Electrocardiogram (ECG), 1820
 open science approach, 1823
 related ethics, 1822
- Air-pollution models, 625, 627, 632
- Airway/ventilator, 1472
- Alanine aminotransferase (ALT), 807
- Alberta Stroke Program early CT score (ASPECTS), 1225, 1512
- Algorithmic bias, 400
- Algorithms, 375–378
- Allergic contact dermatitis (ACD), 1412
- Allergic disorders, 1412
- Allergic march, 1414
- All New Zealand Acute Coronary Syndrome Quality Improvement program (ANZACS-QI), 217
- AlphaFold, 637, 1390
- ALS-FTD, 1692
- ALS functional rating scale—revised, 1692
- Alternative medicine (AM), 1248, 1249
- Alzheimer's disease, 1088, 1693
 diagnosis, 1066
 etiology, 1065
 future research, 1068
 machine learning application in clinical work for, 1065–1068
 prognosis, 1068
 therapy, 1066–1068
- American College of Cardiology (ACC), 696
- American College of Medical Genetics (ACMG), 1090
- American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP), 217, 858
- American Heart Association (AHH), 696
- American Medical Association (AMA), 1223
- American Society of Nephrology, 1450
- Amino acid, 1392
- AmoebaNet, 417
- Amplification and Melting Curve Analysis (AMCA), 1332
- Amyotrophic lateral sclerosis, 1692, 1763
 clinical trial analysis, 1693
 PRO-ACT dataset, 1693–1695
 with machine learning, 1694–1697
- Analytical Engine, 1263
- Analytical epidemiology, 1342
- Analytical Pipeline, 1470, 1471, 1476
- Andrology, 868
 male reproductive potential, prediction of, 868–869
 semen analyses, 869
 sperm retrieval success rates, 869–870
 surgical shunt intervention for priapism management, 870
- Anemia, 581
- Anemia Control Module (ACM), 583
- Anemia management, 581
- Anemia management, artificial intelligence
 artificial neural networks, 1445–1446
 expert systems, 1443–1445
 fuzzy systems, 1447–1450
 pathophysiology and treatment of anaemia in CKD, 1442
 reinforcement learning, 1446–1447
- Anesthesiology, artificial intelligence applications, 1460
 controlling anesthesia delivery, 1455–1456
 event and risk prediction, 1459–1460
 hypotension prediction, 1458–1459
 implications, 1461–1462
 monitor depth of anesthesia, 1455
 perioperative haemodynamic optimization, 1456–1459
- Anomaly detection, 495
- Anonymization, 353, 493
- Anopheles
 biological feature characterization, 1364–1366
 geometric morphometrics, 1362
 vector ecology, 1355
- Anorexia nervosa, 1653
 activity monitoring, 1656
 early diagnosis, 1655
 intake accompanying phenomena, 1657
 psychiatric ED, 1653
 stress reactivity, 1656
- Anterior-segment OCT, 1539
- Antibiotic prescribing, 1337
- Antibiotic susceptibility, 1332
- Antibodies and epitopes, 1392
- Antidepressants, 1568, 1572–1574

- Antimicrobial(s), 1329
 drug design, 347
 selection, 1335
Anti-microbial peptides (AMP), 1392
Antimicrobial resistance (AMR), 1328, 1338
Antimicrobial susceptibility
 approaches, 1332
 determination, 1332
 in-vitro, 1332
 RNAseq data, 1332
Anura™, 1151, 1152
Anxiety disorders, 1568–1576, 1616
Apomediation, 806
Apple Create ML (Apple), 475
Aquamous cell carcinoma (LUSC), 763
Arbitrary arterial sites, 691
Architecture design, 414
Area-under-the-curve (AUC), 525, 572, 589, 697, 859, 1266, 1569
Area under the receiver operating characteristics (AUROC), 928
Arrhythmias
 bradyarrhythmia, 818
 CNN, 819
 diagnosing, 818
 ECGs, 818
 heartbeats, 818
 monitoring, 819
 tachyarrhythmias, 818
Arteriosclerosis, 816
Arteriovenous fistula (AF), 590
Artificial intelligence (AI), 4, 216, 228, 256, 388, 408, 470, 522, 524, 532, 582, 594, 601, 690, 691, 702, 704, 748, 760, 766, 804, 827, 836, 940, 952, 1003, 1008, 1018, 1098, 1171, 1248, 1343, 1347, 1356–1357, 1389, 1391, 1470, 1490, 1491, 1498, 1596, 1622–1626, 1692, 1718, 1727, 1734, 1756, 1778
acupuncture, 1254
ageing biomarker development, 1064–1065
in Alzheimer's disease diagnosis and drug development, 1067
anemia management (*see* Anemia management, artificial intelligence)
ancient tales, 204
in anesthesiology (*see* Anesthesiology, artificial intelligence)
applications, 6–7, 1264
applications in imaging, 1554
applications in ophthalmology, 1521–1535
approaches, 690
approaches in prediction, validation and analysis of stem cells, 1099
autonomous objects, 130
based tools, 1734
in biology, 1062–1063
cancer detection, 1265
cancer treatment, 1268
clinical adoption, 532, 533
clinical applications in ophthalmology, 1554–1564
in clinical immunology (*see* Clinical immunology)
computational methods, 1101–1102
computer, 206–209, 211, 212
computer science, 1264
computing requirements, 1383–1384
concepts and components of, 1521, 1522
COVID-19 (*see* Coronavirus disease (COVID-19), artificial intelligence)
data communication tools, 130
datasets, 388, 975–976
data ownership and sharing, 1543–1544
deep learning, 16–17, 1104
definition, 747, 1264
demand for, 6
for dementia, 1677–1681
in dermatology (*see* Dermatology, artificial intelligence)
diagnosis, 1063–1064
in diagnostic radiology, 463–464
for drug discovery, 1060–1062
education, 1544
for embryo annotation, evaluation, and selection, 1011
for embryo chromosome status (ploidy), 1012
ENT, 987
and evolutionary theory (*see* Evolutionary theory and artificial intelligence)
factor analysis, 693–695
forecasting, 695
futures studies, 1745, 1746
guidelines, 1544–1545
in healthcare, 1059–1065
HER/laboratory data, cancer diagnosis, 1268
history of, 5–6, 745–747
implementation in stem cell progression, 1100
index test, 388
infection biology (*see* Infection biology)
for infertility assessment, diagnosis and treatment, 1011
in interventional radiology (*see* Interventional radiology (IR))
intracerebral hemorrhage management, 1745
intra-operative and robotics applications in orthopaedics, 878–881
invasive procedures, 816
juristic challenges, 133, 134
in kidney pathology (*see* Kidney pathology, artificial intelligence)
Lewy body dementia, 1682
and LHS, 1400–1401
limitations/challenges, 996, 997
machine learning, 1100–1101, 1181
maternal healthcare miscarriage prediction, 1012–1013
in maxillofacial surgery (*see* Maxillofacial surgery)
for medical decisions (*see* Medical decisions)
in medical diagnosis (*see* Medical diagnosis) and medical epistemology, 373–379 and medical ethics, 379–382
in medical imaging, 460–462
medical specialities for education, 333
in medicine, 984, 1063–1065
in Ménière's disease (*see* Ménière's disease (MD))

- Artificial intelligence (AI) (*cont.*)
- and mHealth (*see* mHealth)
 - modalities of clinical utility, 1554
 - modern businesses and technological developments, 130
 - motor and gait impairment detection, 1682–1683
 - multi-agent approach, 1381–1382
 - multilayer perceptron, 17–18
 - myth, 206, 211
 - nanomedicine research, 1172
 - natural language processing, 19
 - non-invasive evaluations, 816
 - nanomedicine research, 1172
 - numbers, 204, 212
 - and nursing education, 753
 - in obstetrics and gynaecology, 334
 - ophthalmology, dermatology, cardiology, gastroenterology, rheumatology, 334
 - in orthopaedic diagnostics, 875–878
 - parameter uncertainty, 1384
 - in Parkinson’s disease, 1681–1686
 - patients’ and physicians acceptance, 1544
 - patient selection, 388
 - pharmacology, 1255
 - philosopher, 204
 - polyp characterization, 978
 - polyp detection and localization, 976–977
 - polyp segmentation, 977–978
 - precision examination and exam practice, 335–336
 - primary healthcare (*see* Primary healthcare)
 - prognosis, 1064
 - programming languages, 7–8
 - psychosomatic therapy, 1254
 - quality assessment tools, 392–393
 - recurrent neural network approach, 1380
 - reference standard, 388
 - in rehabilitation (*see* Rehabilitation)
 - remote diagnostics, 130
 - reporting standards, 389–392
 - research ethics, 1543
 - retinal diseases using fundus photographs, 1524–1527
 - retinal diseases using optical coherence tomographs, 1536–1537
 - in rheumatology (*see* Rheumatology, artificial intelligence)
 - risk assesments, 816
 - SARS-CoV-2, 1382–1383
 - social dilemmas, 131, 132
 - stages of, 7
 - stroke management in subacute phase, 1743–1745
 - surgical performance assessment, 334–335
 - surrogate measurement, 691, 693
 - TCM, 1253
 - thinking machine, 206, 212
 - tools, 528, 1264
 - transient ischemic attack management, 1745
 - unbiased reporting of meta-analysis of improvement in, 332
 - in urology (*see* Urology, artificial intelligence)
 - vaccination, 132
- Artificial intelligence, breast thermography
- biomarker prediction, 1311
 - breast segmentation, 1308
 - CNN based features, 1310
 - hotspot features, 1309
 - malignancy classification, 1309–1310
 - pixel level classification, 1310
 - risk estimation, 1310
 - textural and statistical features, 1309
 - vascular deformations, 1310
 - view labeling, 1308
- Artificial Intelligence (AI), in acute stroke medicine
- applications, 1505
 - black box problem, 1513–1514
 - challenges, 1513–1516
 - clinical outcomes, 1511–1512
 - clot detection, 1510
 - data registries, 1516
 - decision assistance, 1513
 - diagnosis, 1505–1510
 - imaging outcomes, 1510–1511
 - integration, 1512–1513
 - lesion segmentation, 1507–1510
 - model evaluation, 1514–1516
 - outcome prediction, 1510–1512
- Artificial Intelligence Evoked Target Testing (A.R.I.E.T.T.A) project, 1437
- Artificial intelligence in medicine (AIM), 58, 210, 268
- applications, 64–65
 - cardiac arrest (*see* Cardiac arrest)
 - clinical decision support system, 67–71
 - cognitive capabilities, 268
 - difficulties, 276
 - EHR, 271
 - and electronic health records, 61–62
 - financial aspects of, 65
 - history, 59–61
 - Learning Health Systems, 62–63
 - thinking, 270
 - WI lab, 268
- Artificial muscle actuators (AMAs), 900
- Artificial neural network (ANN), 17, 18, 22, 70, 408, 530, 583, 702, 751, 868, 891–893, 910, 1013, 1118, 1206, 1207, 1265, 1331, 1357, 1403, 1404, 1444–1446, 1448, 1450, 1471, 1494–1495, 1597–1599, 1710, 1734, 1736, 1820
- Artificial pancreas (AP), 678, 706
- Artificial reproductive technologies (ART), 869
- Assessment, 906–911
- Assisted reproductive technologies (ART), 1008
- Assisted reproductive technology failure rate, 1010
- Assistive robotic technologies, for dementia, 1679
- Association rule-mining, 305
- Bayesian methods for, 306
 - frequent association rule mining, 305–306
- ASU-Mayo Clinic, 975
- Atanasoff-Berry Computer (ABC), 208
- ATHENA smart process management system, 1800
- Atlas-based techniques, 1721

- Atrial fibrillation (AF), 654–655, 818, 1483
in stroke patients, 1744
- Auditing, 403
- Augmented intelligence, 906
- Augmented reality (AR), 844, 946
- Augment reality-based (AR) systems, 503
- Autism spectrum disorder (ASD), 379
AI applications, 1580, 1581
AI-based research, 1589, 1590
automative/early diagnosis, 1581
brain structural imaging, 1584
confounders, 1590
definition, 1580
drug discovery, 1583
explorative analysis, 1581, 1582
field-based categorization, 1583
genetic liability, 1580
genetics, 1586, 1588
longitudinal, 1582
miscellaneous, 1589
ML/DL, 1591
multi-faceted nature, 1591
PET, 1585, 1586
severity recognition, 1581
subtypes, 1582
teaching interaction, 1583
video/sensor analysis, 1588
- Auto-contouring, 1279
- Autoencoder, 1599–1602
- Autoimmunity, 1436
- Automated breast ultrasound (ABUS), 1296
- Automated decision making, 166, 378, 399
- Automated decision support system (AI-DSS), 656
- Automated decision systems, 1004
- Automated deep learning
AutoML Vision, 482
computer resources, 475, 476
confusion matrices, 481
data preparation, 479
data protection/privacy, 476
large, well curated datasets, 476
limitations, 481, 482
medical imaging, 478
medicine, 478
packaging/deployment models, 483
principles, 477
process, 480
technical expertise, 475
use cases, 483
- Automated diagnostic aids, 533
- Automated dietary monitoring
ambient technology-based, 1648
embedded systems, 1649
long-term monitoring, 1644
processing pipeline, 1645
smart eyeglasses, 1647
smartphone-based, 1648
wearable-based, 1645–1648
- Automated machine learning, 477
- Automated methods, 161
- Automated post processing, of perfusion imaging data, 1739
- Automated surgery, 123–124
- Automated systems, 1336
- Automated televillages, 1241
- Automated thermography, 1312
- Automated time series machine learning, 1348
- Automatic beam selection method, 1279
- Automatic quality control system (AQCS), 654
- Automation, 124, 910, 1281, 1285, 1389, 1392
Ars Magna, 205
definition, 204
figure, 205
for fluid therapy, 1457–1458
Golem, 205
for inotropes infusion, 1458
myth, 204
for vasodilators infusion, 1458
for vasopressors titration, 1458
- AutoML Vision (Google), 478, 1128–1130, 1140
- Autonomous intraoperative robots, 880
- Autonomous robotic surgery, 336
- Autonomous systems, 1456, 1458
- AutoPiLoT AI-system, 1166
- Autopsy, 1772
- Autoregressive model, 583
- Auto-segmentation, 1279
- AVATAR therapy, 1604
- Avicenna, 1388
- B**
- Babylonian period, 1388
- Backpropagation, 460, 531
- Bacteria, 1371–1373
- Bacteriorhodopsin, 662
- Bag of Keypoints (BoK), 792, 799
- Bag of Keypoints Classifier*, 797, 798
- Bag-of-visual Words (BoW), 847
- Ballistocardiogram (BCG), 692
- Barrett's esophagus (BE)
AI, quality assessment, 952, 959, 960
CNN, 953, 957
definition, 952, 953
endomicroscopy, 958, 959
endoscopists, 961
NBI/VLE, 961
NBI, 957
visual identification, 952
white light endoscopy, 953
- Barrier reduction intervention (BRI), 656
- Basal insulin, 706
- Base-learner model, 409
- Baycrest Centre for Aging and Brain Health Innovation (CABHI), 1616
- Bayesian diagnosis, 189–191
- Bayesian inference, 189
- Bayesian models, for decision support, 169–171
- Bayesian networks (BNs), 189–191, 1335
for CDSS, 170–171

- Bayesian rule mining, 306
 BBMAP and BBLL techniques, 799
 B-cell epitopes, 1390
 Behavioral health, 1614
 Behavioral intervention technologies (BITs), 1617
 Benchmark datasets, 402
 Beneficence, 140–141
 Benefits, 1611–1613
 Berg Balance Scale (BBS), 1801
 Bernstein-Vazirani algorithm, 440
Best First Search (BF), 790
 BG control strategies, 706
 BG prediction, 706
 Bias, 119, 122, 124–126, 232–234, 372, 373, 376–379, 382
 Bidirectional encoder representations from transformers (BERT) model, 640, 1090
 Big data, 294, 299, 463, 1276, 1278, 1280, 1283–1285, 1343, 1349, 1399
 in radiology, 463
 for toxicology interpretations, 1497–1498
 Big data-based AI technologies, 289
 Binary object morphology, 530
 Biochemists, 1392
 Biomarker(s), 119–121, 374, 1631
 identification, 1404
 Biomarker discovery
 allergic diseases, 1415
 ML approaches, 1415
 Biomarker identification, 1404
 Biomedical sciences, 1389
 Biomedical technology, 294
 human perfection and security implications, 295–298
 Biotrauma, 1473
 BiteScope, 1363
 Black box, 1005
 approach, 694
 model, 507, 508
 systems, 1764
 Bloch sphere, 433
 Block-based approach, 799
 Blockchain, 359, 360
 Blood antigen detection, 1435
 Blood Examination Records (BERs), 271
 Blood glucose (BG), 702
 Blood pressure (BP), 690, 691, 695, 696
 general health assessment, 1145
 general health monitoring, 1145
 hypertension management, 1144–1145
 Blood stasis syndrome, 1256
 Blood tests, 723
 Blood transcriptomics
 clinical applications, 1119
 clinics/methods/features/pitfalls, 1112, 1113
 data governance, 1111
 data security/federated learning, 1120
 DGE, 1114
 DNNs, 1118
 high-throughput technologies, 1110
 LASSO, 1117, 1118
 random forest, 1115
 rich feature space, 1120
 supervised learning, 1114
 SVMs, 1116, 1117
 unsupervised learning, 1115
 Blue Laser Imaging (BLI), 942
 Body surface area (BSA), 554
 Bone and mineral disorders, AI applications
 bone metabolism, 680
 fracture identification, 679
 fracture risk assessment, 680
 opportunistic screening of osteoporosis and sarcopenia, 679–680
 Bone marrow haematopoietic stem cell transplantation, 1435
 Bone mineral density (BMD), 781
 Boolean logic gates, 429
 Boosted regression tree (BRT), 626
 Borderline personality disorder (BPD), 377
 Bose-Einstein condensates, 425
 BP measurement methods, 690
 Bradycardia, 818, 1480, 1481
 Brain age gap (BAG), 1064
 Brain structural imaging, 1584
 Brain tumors
 atlas-based techniques, 1721
 BraTS, 1726, 1727
 characteristics, 1725, 1726
 classification, 1726, 1727
 data sets, 1722
 detection, 1720, 1722
 DICE scores, 1724
 heterogeneity, 1721
 hybrid approaches, 1722
 medical-image segmentation pipelines, 1721
 MRI (*see* Magnetic resonance imaging (MRI))
 multi-modal processing, 1722
 qualitative analysis, 1725
 quantitative analysis, 1724
 segmentation, 1720, 1722
 statistical analysis, 1725
 Brain tumor segmentation (BRATS), 504
 Brain Tumor Segmentation Challenge (BraTS), 1722
 Breast cancer detection, 1288
 BCDR dataset, 1289
 DDSM dataset, 1289
 in deep learning methods, 1291
 future technologies, 1295–1297
 INbreast dataset, 1289
 MIAS dataset, 1289
 OMI-DB dataset, 1289
 performance metrics, 1294
 Breast Cancer Detection Demonstration Project (BCDDP), 1306
 Breast Cancer Digital Repository (BCDR), 1289
 Breast cancer risk estimation, 1310
 Breast infrared assessment system (BIRAS), 1310
 Breast infrared thermography

- artificial intelligence for, 1307–1311
breast segmentation, 1308
challenges with manual interpretation of, 1306–1307
description, 1305
imaging protocols, 1305
malignancy classification, 1309–1310
view labeling, 1308
- Breast pathology, 525
- Breast segmentation, 1308
- Breath analysis, 765, 1210
- Breath detection algorithm (BDA), 1803
- British Society of Rehabilitation Medicine (BSRM), 1812
- Bronchoscopic inspection, 763
- Brownian motion, 432
- Bruch's membrane opening-minimum rim width (BMO-MRW) parameter, 1558
- Bulbar stratification, 1692
- C**
- CAD4TB, 66
- Calibration, 692
- Calinski-Harabasz score, 1133, 1134
- Camelyon dataset, 528
- Cancer, 343–345, 525
disease progression, 1433
drug discovery, 1175
immunotherapy, 1433
registries, 1343
- Cancer detection
clinical images, 1266
pathological images, 1267
radiological images, 1265–1267
- Cancer treatment
clinical decision making, 1268, 1269
radiation oncology, 1269, 1270
surgery, 1269
- Capitalism, 381–382
- Capsule endoscopy (CE), 940
- Carbapenemase Producing Enterobacteriaceae (CPE), 1332
- Carbon-based systems, 425
- Cardiac arrest, 1484
active intervention and follow-up, 1483
AI, 1482, 1484
clinical treatment, 1481–1482
continuous monitoring and subsequent interventions, 1483–1484
definition, 1480
early recognition/detection of high-risk patients, 1482–1483
SCA, 1480
ventricular arrhythmia, 1480, 1481
- Cardiac computed tomography (CT), 814
- Cardiac rehabilitation (CR), 1236
- Cardiopulmonary resuscitation (CPR), 1472
- Cardiovascular diseases (CVD), 272, 690, 698
- Cardiovascular disorders
AI programs, 814
arrhythmias, 815
diagnosing, 814
heatmap, 820
recognition capability, 819
stenosis, 816
video image data, 815
- Cardiovascular magnetic resonance imaging (CMR), 814, 1482
- Cardiovascular medicine, 1236–1237
- Care, 372, 377, 381
- Care-O-Bot, 1679
- Carpal tunnel syndrome, 1762
- Cartilage segmentation, 877
- Case-based reasoning (CBR), 68–69, 702, 1335, 1336
- Case-control design, 563
- Catalogue of Somatic Mutations in Cancer (COSMIC), 1269
- CATAPULT, 1076
- Catastrophic forgetting, 413
- Catheter-based interventions, 814
- Causability, 506
- Causal biases, 172
- Causal Cognitive Architecture 1 (CCA1), 26, 28, 1603
- Causal diagnostics, 290
- Causal discovery of diseases, 1695
- Causal inference, 693, 694
- Causality(ies), 171–173, 1639
- Causal-Probabilistic-Networks, 1334, 1335
- Causal tendency-entropy ratio (CTER) measure, 1697
- Cell-based immunity, 1388
- Cell-based search, 416
- Centers for Disease Control (CDC), 597
- Central line-associated bloodstream infections (CLABSIs), 219
- Central nervous system (CNS) tumors, 1665
- Centres for Disease Control and Prevention (CDC), 222
- Centres for Medicare and Medicaid Services (CMS), 222
- CENTRIST, 1794
- Cerebral blood flow (CBF), 1585
- Cervical cancer, *see* Pap smear test
- Cervical spondylotic myelopathy (CSM), 1763
- Cervical VEMP (cVEMP), 1708
- Chaos-assisted tunnelling, 435
- Chase's contract intelligence platform (COiN), 6
- Chatbots, 118, 1238
24/7 access for 430,000 people, 1613
AI-assisted, 1277, 1283, 1284
AI mental health support for African young mothers, 1614–1616
basic concepts, 1283
cloud-based development platforms, 1283
comorbidities in childhood obesity, pre-diabetes, and mental health struggles, 1617–1618
economic impact, 1611–1613
healthcare costs, 1283
mental health, 1610
radiotherapy patients, 1284

- Chatbots (*cont.*)
 in real world applications, 1613–1618
 support for caregivers and patients, 1616
 value of mental health, 1610–1611
 workflow, 1284
 working, 1611
- Chattecx balance system (CBS), 1801
- Chemical genomics-based approach, 638
- Chemical sensors, 1204, 1205, 1210, 1211, 1213
- Chemoreceptors, 1204
- Chest x-ray (CXR), 761, 1400
- CheXNet, 375, 1238
- Childhood cataracts, 655
- Childhood medulloblastoma, 1665
- Chinese acupuncture expert system (CAES), 1254
- Chlamydia pneumoniae*, 1334
- Chrominance method, 1148
- Chronic disease self-management, 276
- Chronic kidney disease (CKD), 580, 1442
- Chronic obstructive pulmonary disease (COPD), 763, 1237
- Church-Turing thesis, 426
- Civil 'drones, 134
- CKD-Mineral Bone Disorder (CKD-MBD), 589
- CKD secondary anemia (CKD-anemia), 581
- Clarifai Train (Clarifai), 475
- Classical computation, 429
- Classical machine learning, 1721
- Classical regression methods, 1343
- Classic wrapper method, 1416
- Classification accuracy, 1571
- Classification and Regression Trees (CART), 165, 626
- Classification or regression techniques, 705
- Classification problems, 11
- Classifiers, 77, 260
- Classifiers and discriminant functions, 791
 bagging, 791
 Bayes network, 791
 MAP decision rule, 791
 multilayer perceptron, 791
Naïve Bayes (NB), 791
 Random forests, 791
- Clinical data
 definition, 241
 diagnostic-related information, 243
 EHRs, 241, 242
 omics, 242, 243
 precision health, 241
- Clinical Data Management (CDM), 241
- Clinical decision making, 161, 173, 182, 196, 1337
 causality, 171–173
- Clinical decision support (CDS), 737, 1048, 1250, 1402, 1450
- Clinical decision support systems (CDSS), 67–71, 160, 170, 171, 264, 1329, 1400
- Clinical diagnosis, 1248
 AM, 1252
 TCM, 1256
- Clinical epidemiology, 1343, 1349
- Clinical evaluation of AIM
 blindspots, esophagogastroduodenoscopy, 654
 childhood cataracts, 655
 colonic adenoma, endoscopy, 647–654
 insulin dose, 656
 intraoperative hypotension, 655
 mental health risk assessment, 655–656
 monitoring drug adherence, 656
 new reporting guidelines, 657–658
 paroxysmal atrial fibrillation, 654–655
 randomized controlled trials, 647–656
 robust clinical evaluation, 646–647
- Clinical forensic medicine, 1783
- Clinical immunology, 1398
 biomarker identification in, 1404
 candidate selection for clinical trials and patient identification efforts, 1402–1403
 CDS, 1402
 cytometric analysis, 1405
 disease diagnosis, 1401–1402
 ethical implications in AI, 1405
 microbiome analysis, 1404–1405
- Clinical Interpretations of Variants in Cancer (CIViC), 1090
- Clinical neurophysiology, 1754, 1764
- Clinical neurosciences, 1754
- Clinical practice guidelines, 256, 264
- Clinical quality language (CQL), 70
- Clinical reasoning, 270, 274
 disease diagnosis, 281
 knowledge representation, 280
- Clinical rules, 1334
- Clinical Study Reports (CRS), 262
- Clinical text in the EMR in Chinese (CEMR), 271
- Clinical therapies, 1253
- Clinical toxicology, 1488
- Clinical trials, 636, 638, 640, 728, 1392
- Clinic-based monitoring, 1152
- Clinic effectiveness, 727
- Closed-loop control, 1446, 1448
- Cluster analysis, 1251
- Clustered regularly interspaced short palindromic repeat (CRISPR), 1021
- Clustering, 581, 588, 589, 1570
 algorithms, 1494
 problems, 11
- CNN-based model, 279
- Coaching, 1615
- Cobb angle, 882
- Co-design, 1337
- Cognitive architecture, 268–270
 cerebral cortex, 269
 primary stage, 269
 symbolic-connectionism, 270–272
- Cognitive behavioral therapy (CBT), 726, 1573, 1575
- Cognitive biases, 185
- Cognitive errors, 185
- Collaboration, 364, 365
- Collider bias, 171

- Colorectal cancer (CRC), 968
Colorectal polyps, deep learning methods, 970
Combat diseases, 1388
Combinatorial optimization methods, 168–169
Combined annotation-dependent depletion (CADD), 1089
Community, 367
 physiotherapists, 1803
Comorbidities, 1374, 1632
Complementary and alternative medicine (CAM), 1248
Component correction (CC), 1212
Compounds, 636–638
Comprehensive monitoring, 1152
Comprehensive stroke centers, 1736
Computational analysis, 1389
Computational approaches in biological system, 1102
Computational biases, 378–379
Computational cell kinetics model, 1101
Computational fluid dynamics (CFD), 625
Computational toxicology, 1497
 big data for toxicology interpretations, 1497–1498
 PBPK modelling, 1498–1499
Computed tomography (CT), 503, 762, 1265, 1278, 1584
Computer-aided design (CAM), 827
Computer-aided detection (CADe), 647, 943
Computer-Aided Detection for Tuberculosis (CAD4TB), 620
Computer-aided diagnosis (CAD), 847, 920, 943
Computer-aided drug discovery (CADD), 1061
Computer-aided image analysis, 524
Computer-aided sperm analysis (CASA) systems, 1004
Computer-assisted detection (CADe), 942, 968
Computer-assisted diagnosis (CADx) systems, 942
Computer-assisted navigation (CAN) systems, 878
Computer-assisted sperm analysis (CASA), 869
Computer-based analysis, 1664
Computer-based image analysis, 552
Computer-controlled collimator, 1282
Computerized lung sound analysis, 766
Computerized tomography (CT), 827, 1032
Computer prescribing order entry (CPOE)
 systems, 1329
Computer-supported decision making, 1337
Computer tomography (CT), 1285, 1373
Computer vision (CV), 122–123, 859, 894–895, 1264
 in infection biology (*see* Infection biology)
 and supervised learning, 857
 technical aspects, 857
 and unsupervised learning, 857–858
Computing Machinery and Intelligence, 1264
Conditional probability, 189
Conditional restricted Boltzmann machines, in
 Alzheimer’s disease, 1679
Cone-beam computed tomography (CBCT), 907–909, 911
Cone-bean CT, 1280
Confocal Laser Endomicroscopy (pCLE), 846
Confounding bias, 171
Confounding Index (CI), 1590
Congenital heart disease (CHD), 1032
Consent, 352, 353, 355, 357, 359, 360
Consolidated Standards of Reporting Trials (CONSORT), 389, 657
CONSORT-AI, 532
Constraint-induced movement therapy, 1802
Consumer devices, 691
Contactless approaches, 693
Content-based microscopic image analysis (CBMIA), 1102
Contestability
 bias, 232–234
 decisional role, 235–236
 dimensions of, 231–233
 issues, 236–237
 performance, 234–235
 personal health data, 232
 right to, 229–231
Context awareness
 context types and recognition path, 1191–1192
 definition and theory, 1190–1191
 exposome, 1188
 pattern spotting, 1192
Continental network model, 618
Continual learning, 412–413
Continuous glucose monitoring (CGM), 656, 702, 1035
Continuous glucose monitoring system (CGMS), 678
Continuous imaging protocols, 1305
Continuous monitoring, 403
Continuous prediction, 563
Continuous wavelet transform, 1763
Convalescent plasma (CP), 1403
Conventional breast imaging modalities
 challenges with, 1304
 magnetic resonance imaging, 1304
 mammography, 1303
 ultrasound, 1303–1304
Conventional IVF, 1010
Conventional methods, 707
Conventional randomized clinical trial approaches, 693
Conventional sparse classifiers, 799
Conventional SRC method, 797
Convolution, 78
Convolutional network, 504
Convolutional neural nets, in dementia, 1681
Convolutional neural network (CNN), 18, 42, 77, 94, 248, 335, 450, 462, 530, 552–554, 606–607, 638, 775, 781, 827, 910, 924, 942, 986, 1090, 1240, 1252, 1331, 1357, 1371–1373, 1375, 1390, 1398, 1405, 1496, 1584, 1706, 1709, 1710, 1712, 1713
 backward pass, 80
 convolution, 78
 cross-correlation, 78
 deep learning (*see* Deep learning)
 forward pass, 79
 layered structure, 78
 network topologies, 85–88
 representation learning, 83
 transfer learning, 85
 unsupervised learning, 81

- Convolutional rObust pRincipal cOmpoNent Analysis (CORONA), 49
- Convolution neural network (CNN), 760
- Convolution-superposition algorithm, 1279
- Cooperative Neural Networks, 1131
- Coregistering coordinate systems, 828
- Coronary angiography (CAG), 816
- Coronary arteries, 816
- Coronavirus disease (COVID-19), artificial intelligence clinician demographics and image features, 514–516
- COVID-19 diagnosis, 516–517
 - education, 514
 - medical imaging, 512–514
- Coronavirus infectious disease 19 (COVID-19), 342, 343, 348, 1193, 1222, 1374, 1403–1404
- haemo-virology, 1434
 - infection detection, 809
 - pandemic, 533, 1345, 1350, 1380
- Coronavirus Protection Index (CPI), 1382
- Corpus construction, 277, 278
- Correlation-based Feature Selection (CFS), 790
- Cost of customer acquisition (CoCA), 605
- Counterfactual framework, 694
- Crete-Roffet Blur Metric (CRBM), 846
- CRISPR/Cas9, 298
- Criteria2Query, 1235
- Critical Assessment of Structure Prediction (CASP), 666
- Critical care, 1470, 1472
- AI/ML applications, 1474
 - AI, 1475
 - clinical trials, 1472
 - ML, 1473
- Critical view of safety, 856
- Cross-sectional data, 695
- CT angiography (CTA), stroke assessment, 1506
- CT perfusion imaging (CTP), stroke assessment, 1506
- CT pulmonary angiography, 762
- Cuff-based BP monitors, 692
- Cuffless, 690, 691
- CURATE.AI-guided treatment, 1180
- CVC-EndoSceneStill, 974
- CVC-VideoClinicDB, 975
- Cyber-physical systems, 1819
- Cybersecurity, 356
- Cystic fibrosis (CF), 1405
- Cystoscopy, 866
- Cytokines, 1388
- Cytometric analysis, 1405
- Cytotoxic T Lymphocytes (CTL) epitopes, 1390
- D**
- Daily practice, 1778
- Darwinian theory, 343
- Data, 366, 1344
- analytics, 1136–1140
 - augmentation, 970, 1723
 - digitization, 1343
 - in medical imaging, 463
 - protection laws, 617, 618
- science, 343, 347, 348, 1398, 1399, 1401, 1403, 1405, 1406
- science methods, 263
- security and cybersecurity, 300
- Data-centred practice, 372
- Data-dependence, 1471
- Data extraction, 260–262
- time-consuming, 262
- Data-mining, 1285
- Data processing, 1348
- dimensionality reduction, 245
 - missing value imputation, 244
 - omics processing, 245, 246
- Datasets, 375
- Dataset shift, 403
- Data sharing
- public's view on, 355–356
 - and regulations, 353–354
- Data Sharing Project (DSP), 222
- Data tier, 272
- Data-utopianism, 374–375
- 3D Avatar, 1804
- Davies-Bouldin score, 1134
- daVinci system, 826
- 3D convolutional neural networks model, 1061
- De-anonymized raw data, 1337
- Death, 1768, 1772
- Decision support platforms, 1241
- Decision support systems (DSS), 702, 704–706, 708
- Decision tree, 12, 274, 586
- Decision-tree-based methods, 694
- Decision Tree Classifiers, 1334
- Deep aging clocks, 1159–1160
- Deep Autoencoders (DAE), 1474
- Deep belief neural network (deepBN), 248
- Deep brain stimulation (DBS), 1685
- DeepChem, 1391
- Deep convolutional neural network (CNN), 76, 1012
- Deep convolution neural network based algorithm, 1741
- Deep feature selection (DFS) model, 1064
- DeepGestalt, 1091, 1093
- Deep iterative registration, 829
- Deep learning (DL), 5, 16–17, 22, 49–52, 76, 88, 121, 167–168, 210, 212, 268, 269, 335, 336, 408, 462, 489, 504, 512, 540, 542, 584, 677, 679, 680, 747, 774–776, 779–781, 856–858, 891, 894, 898–900, 902, 906–908, 942, 952, 986, 1060, 1104, 1158, 1224, 1225, 1252, 1265, 1357, 1370, 1375, 1390, 1392, 1471, 1491, 1496, 1498, 1520, 1581, 1622, 1624, 1626, 1664, 1680, 1681, 1706, 1712, 1714, 1721, 1756, 1758
- age-related macular degeneration, 1561–1564
 - algorithms, 524, 525, 877, 878
 - algorithms for retinal diseases, 1535–1539
- AMD, 1529
- approaches, 194, 195, 1737
- bounding-box, 973
- cell segmentation, 1669
- classification, 88, 1669

- classification network, 971
cycleGANs, 90
datasets, 975–976
definition, 523, 530
detection, 88
diabetic retinopathy, 1521–1528, 1555–1557
for electromyography, 1762–1763
electronic health records, 1540
in epilepsy treatment, 1760
factors, 531
feature extractor, 971
GANs, 90
glaucoma, fundus photographs, 1528–1529
glaucoma, OCT, 1539
glaucoma, 1556–1560
image quality assessment, 1540
image reconstruction, 88
image registration, 90
image restoration, 90
infantile facial video recording, 1540
for investigating epilepsy, 1758–1759
neural radiance fields, 90
PACS, 90
papilledema and optic disc abnormalities, 1530
patch-based, 973
phases, 531
polyp characterization, 978
polyp detection and localization, 976–977
polyp segmentation, 977–978
retinopathy of prematurity, 1529
revolution, 23
segmentation, 88
semantic segmentation, 973
speech recognition and translation, 88
super-resolution, 89
systemic diseases, 1530–1534
visual fields, 1540
Deep learning-based sequence analyser (DeepSEA), 1090
Deep-learning funduscopic atherosclerosis score (DL-FAS), 1531
Deep-learning health systems, 265
Deep Learning Important FeaTures (DeepLIFT), 249
Deep learning methods
 in CAD system, 1296
 DDSM dataset, 1292
 state-of-the-art, 1291, 1292
 You Only Look Once (YOLO), 1293
Deep machine learning (DML), 751
DeepMind, 354, 356
 artificial intelligence system, 60
 platform, 64
Deep neural network (DNN), 23, 507–508, 666, 692, 697, 760, 775, 792, 797, 857, 883, 930, 986, 1118, 1160, 1496, 1599
data-set topologies, 452, 453
definition, 450
hierarchical feature extraction, 450, 451
single-layer perceptron, 450
Universal Approximation Theorem, 450
Deep reinforcement learning (DRL), 133, 1474
DeepSynergy, 1092
Deep-Vac-Pred, 1390
DeepVS, 1061
De-identification methods, 148
Delphi method, 1242
Delusion, 1596
Dementia, 1676
 artificial intelligence for, 1676–1681
 ethical and social implications, 1686
 in-vivo detection and management, 1686
 patient video monitoring and analysis, 1679
 related electroencephalographic analysis, 1680
De novo drug design, 1062
Dentistry, AIM, 915
 applications, 913–914
 assessment, 907–911
 diagnosis, 911–912
 outcomes prediction, 912–913
 treatment planning, 912
Depression, 1610, 1614, 1616
Depression and anxiety, AI in
 diagnosis, 1571–1572
 risk, 1571
 suicidality, 1574
 treatment outcomes, 1572–1574
Depth-from-intensity technique, 960
Dermatological image analysis, 1266
Dermatology, artificial intelligence, 730, 1240
 dermatologist attitude, 556
 eczema, 555–556
 ethnic variations, 557–558
 limitation, 556–557
 psoriasis, 554–555
 skin cancer, 552–554
 skin disorders, 556
 teledermatology, 557
Descriptive epidemiology, 1342
Detection, 922
Determination of causes of death, 1782
Deutsch-Jozsa algorithm, 440
Developing countries, 617
Diabetes
 AI based tools, 705
 behaviors changes, 706
 BG levels, 703, 707
 cardiovascular complication, 707
 comorbidities, 707
 diabetic patients, 707
 diagnosis, 705
 growth, 706
 health care, 708
 lifestyle, 706
 management, 702, 705
 misdiagnosis, 705
 normoglycemic range, 702
 occurrence, 703
 prediabetes, 703, 705
 prognosis, 705

- Diabetes mellitus, AI applications
 continuous glucose monitoring and closed-loop
 artificial pancreas system, 678
 prediction of diabetic complications, 677–678
 retinopathy detection, 677
 therapeutic lifestyle modification, 678–679
- Diabetes nutrition medical record (DNMR), 271, 276
- Diabetic comorbidities, 707
- Diabetic ketoacidosis, 705
- Diabetic nephropathy (DN), 541
- Diabetic retinopathy (DR), 211, 677, 707, 1224,
 1521–1528, 1555–1557
 diagnosis system, 606
 fundus photographs, 1520–1524
- Diagnosis, 274, 906, 907, 910–912
 algorithms, 184
- Diagnostic and Statistical Manual of Mental Disorders
 Fifth Edition (DSM-5), 1596
- Diagnostic haemopathology using artificial intelligence
 neural networks, 1429
 polycythaemia vera screening, 1430
 pyruvate kinase deficiency, 1430
 thalassaemia screening, 1429
 WBC classification, 1429
- Diagnostic radiology (DR), 460, 463–464
- Dialysis, 1442, 1443, 1445, 1446, 1448–1450
- Diastolic BP, 691
- Dice index, 1309
- Dice similarity coefficient (DSC), 925, 1515
- Differential gene expression (DGE), 1114
- Differential privacy (DP)
 applications, 152
 challenges, 152
 definition, 151
 implementation, 151
 properties, 151
 sensitivity/privacy budget, 151
- Diffraction limit, 1370
- Digital biomarkers, 121, 1645, 1647
 classification algorithms, 1650
 eating timing detection, 1647
 food amount, 1652
 food type recognition, 1650
 segmentation, 1650
 triggers and stressors, 1653
- Digital breast tomosynthesis (DBT), 1296
- Digital data, 1346
- Digital Database for Screening Mammography
 (DDSM), 1289
- Digital equipment, 1476
- Digital health, 1188, 1221
 digital biomarkers, 1192
 digital twins, 1195
- Digital indicators, 1347
- Digitalization, 488
 of the consumer, 134
- Digital microbiology, 1373
- Digital morphological analysis, 1430
- Digital pathology, 522–525, 528, 530, 532, 533
- Digital signal filters, 1149
- Digital slides, 525, 528, 530, 531
- Digital therapeutics, 125, 679
- Digital Twins (DTs), 250
- Digitization, 530
- Direct dose optimization systems, 587
- Directed acyclic graph (DAG), 170, 190
- Direct numerical simulations (DNS), 625
- Direct oral anticoagulants (DOACs), 656
- Disaggregated data, 402
- Discrete Fourier and Cosine Transforms, 788
- Discrete imaging protocol, 1305
- Discrete wavelet frames, 787
- Discrete wavelet transform (DWT), 1211
- Discrete wavelet transform and long short-term memory
 (DWTLSTM), 1211
- Disease-centered treatment, 276
- Disease endotyping, 1417
- Disease module, 1075, 1076, 1078
 Largest Connected Component (LCC) Module, 1078
- Disease of anatomical entity, 1083
- Disease ontology (DO), 1081–1083
- Disease phenotyping, 1417
- Disease risk prediction, 282–284, 1414
- DisGeNET database, 1081, 1084, 1603
- Distribution, 638–639
- Dizziness, 1706, 1707
- DNA nanotechnologies, 426, 1392
- Document-term matrix, 19
- Domain adaption, 1723
- Domain regularized component analysis
 (DRCA), 1213
- Donor-recipient body weight (DRBW), 1032
- Dose distribution index (DDI), 1276
- Dose-volume histograms (DVHs), 1280
- Double screening, 260
- 3D proteins structures
 protein folding, 662–663
 secondary structure prediction, 663
- Dreyfus model, 831
- Drug delivery systems and formulation, 1175
- Drug design, 347
- Drug development, 728
- Drug discovery, 1391, 1601, 1602
 deep neural networks (DNNs), 666
 drug repurposing, 666
 generative adversarial networks, 667–669
 inputting molecular structures, 669
- Drug metabolism, 639
- Drug repurposing, 666, 1075, 1076
- Drug resistance, 346–347
- Drug therapy, 1158
- Dual-energy X-ray absorptiometry (DXA), 679, 786
- D-Wave quantum system, 441
- DWI-FLAIR mismatch paradigm, 1737
- Dx-DR algorithm for diabetic retinopathy screening, 1555
- Dynamic contrast-enhanced MRI (DCE-MRI), 1726
- Dynamic Task Prioritization (DTP), 415
- Dynamic tunnelling, 435

E

- Early disease risk assessment (eDRAM), 776
EasiCSDeep, 1763
EasyScan Go (Motic digital[®]), 1361
Echocardiography, 814–817, 1482, 1483
eClinics, 1242
ECG-FLM, 1762
Ecological momentary assessment (EMA), 1575
Economic studies, 65
Eczema, 555–556
Edge Histogram, 790
Education, 753
Efficiency of AI models, 413
eHealth, 1243
EHR History-based prediction using Attention Network (EHAN), 249
Electrocardiogram (ECG), 691, 814, 815, 1482–1484
Electroencephalographic diagnosis, in Parkinson's Disease, 1683
Electroencephalographic patient data, 1602
Electroencephalography (EEG), 1269, 1581, 1585, 1754, 1758
 machine learning, 1759–1760
 montage-based system, 1759
 practical machine learning, 1763
Electromyography, 1763
Electronic cohorts/e-cohorts, 1347
Electronic health records (EHR), 61–62, 119, 217, 240, 242, 263, 270, 271, 408, 608, 616, 618, 704, 705, 713, 715, 724, 728, 737, 774, 776–779, 1030, 1090, 1223, 1265, 1329, 1344, 1345, 1398–1402, 1483, 1540
 medical data, 284
Electronic medical records (EMRs), 271, 294, 571, 573, 1238, 1252, 1443
Electronic noses, for medical diagnostics, 1205–1209
 denoising algorithms, 1211
 sensor drift compensation, 1211–1212
Electronic olfaction, 1204, 1206, 1207, 1211, 1213
Electronic patient records (EPRs), 456
Electron microscopy, 1370, 1372
ELIZA, 1610
Elizzbot, 1616
Embedded method, 245
Embolic Stroke of Undetermined Source (ESUS), 819, 1744
Emergency care, for haemorrhagic stroke, 1504
Emergency Department (ED), 217
Emerging infections, 1370
Employee Assistance Program (EAP), 1611
EMR quality control, 275
Enclave, 154
Endocrinology, AI applications, 682
 bone and mineral disorders, 679–681
 diabetes mellitus, 675–679
 pituitary and adrenal disorders, 682
 thyroid disorders, 681–682
Endocrinology, 1239–1240
Endolymphatic hydrops (EH), 1707, 1709, 1713

Endomicroscopy, 958

- Endoscopy
 anatomical districts, 940
 applications, AI, 942
 AR/VR, 946
 challenges, 947
 definition, 940
 detection/diagnosis, 942, 943
 DL, 947
 informative frame selection, 944
 mosaicking/surface reconstruction, 945
 procedure, 941
 videos, 830
Endourology, 867–868
Endovascular clot retrieval (ECR), 1504
Endovascular recanalization, 1736, 1743
End-Stage Renal Disease (ESRD), 580, 588
End-to-end and semantic segmentation models, 977
End-to-end methods, 973
End-to-end systems for malaria diagnosis, 1361
Energy-based Generative Adversarial Network (EBGAN), 1130
Enhancing Quality and Transparency of Health Research (EQUATOR), 657
Entropy, 788
Entscheidungsproblem, 426
Environmental data, 1343
Enzyme linked immunosorbent assay (ELISA), 1403
Epidemiology
 causal inference, 1349
 challenges, 1348
 data, 1344
 data and interpretation, 1350
 definition, 1342
 numerical methods, 1343
 origin, 1342
 protocolized/standardized production, 1343
 risk factors, 1348
 use cases, 1344
Epigenetic clock, 1159
Epilepsy
 description, 1755
 machine learning and deep learning, 1757–1759
 surgery, 1761
 treatment, 1760–1761
Epitopes, 1388
Erythrocyte sedimentation rate (ESR), 1239
Erythropoiesis process, 582
Erythropoiesis stimulating agent (ESA), 581, 1442–1450
Erythropoietin, 1442, 1445–1447, 1450
 synthesis, 581
Esophageal adenocarcinoma (EAC), 952, 953
Esophageal cancer, 952
Esophagogastroduodenoscopy (EGD), 654
ETIS-Larib, 975
European Centre for Disease Prevention and Control (ECDC), 346
Evaluated an AI detection system (ENDOANGEL), 654
Evaluative epidemiology, 1342

- Evidence-based diagnosis, 1329
 Evidence based individualized treatment pathways, 125
 Evidence-based medicine (EBM), 71, 256, 604, 713, 1058, 1337
 criticisms, 256
 definition, 256
 mHealth, 1235
 vision, 257
 Evidence-based policy (EBP), 596
 Evidence synthesis, 262
 Evolutionary coupling (EC), 248
 Evolutionary theory and artificial intelligence
 infectious disease, 345–348
 mathematical oncology, 343–345
 eWardrounds, 1242
 Excretion, 639
 Expert(s), 1770
 knowledge, 1250
 system, 596, 702, 704, 705, 1443–1445, 1597, 1598
 Expertise, 1778
 Explainability, 236, 291, 380, 508, 1471
 Explainable artificial intelligence (XAI), 249, 505–506, 690, 693, 694
 Explainable ML algorithm, 1269
 Exploratory data science, 373
 Extended Kalman Filter (EKF), 842
 Extracorporeal shock wave lithotripsy (ESWL), 867
 Extrasystole, 818
 Extreme gradient boosting (XGB), 572
 Extremely Randomized Trees (ERTs), 586
- F**
- Facebook, 1347
 Facebook/Cambridge Analytica, 356
 Facial recognition technology, 606
 Factor analysis (FA), 1474
 Fairness, 508
 Familial hypercholesterolemia (FH), 220
 Familial stratification, 1692
 Faster R-CNN model, 1293
 Feasibility theory, 426
 Feature(s), 1568, 1572
 engineering, 573
 extraction, 1205
 reduction, 1668
 relevance, 507
 selection techniques, 245
 Federated computing, 360
 Federated learning (FL), 1161, 1345, 1724
 applications, 150
 attacks, 150
 challenges, 149
 definition, 149
 DP, 151
 systems, 367
 technical framework, 149
 Feed-forward artificial neural network, 1429
 Feed Forward Neural Networks (FFNNs), 583
 Fermions, 425
 Fertility, 1004
 Few Shot Learning, 411–413
 Filter approach, 245
 Findable, Accessible, Interoperable, Reusable (FAIR), 262
 Fine Needle Aspiration Cytological (FNAC), 1240
 Fine needle aspiration (FNA), 990
 First generation (1G) mobile networks, 1230
 Fitzpatrick scale, 1150
 Fixed Reference T-distributed Stochastic Neighbors (FR-t-SNE) method, 848
 Flexible spectral Imaging Color Enhancement (FICE), 942
 FloReMI, 1434
 FlowSight® imaging cytometry platform, 1433
 Fluid Attenuation Inversion Recovery (FLAIR), 1718
 Fluid therapy, 1456–1458
 Focused Assessment with Sonography for Trauma (FAST), 846
 Food and Drug Administration (FDA), 676, 860, 931, 1076, 1443
 Food and Drug Admission, 219
 Forced oscillation test (FOT), 763, 764
 Forensic medicine, 1768, 1778, 1783–1786
 artificial intelligence and clinical, 1771
 artificial intelligence and forensic pathology, 1772
 data, information and evidence, 1770
 direct practice of, 1772
 evidence and individualization in, 1769–1771
 medical expertise and artificial intelligence, 1772
 personalised medicine, 1770–1771
 potential trends and future challenges, 1773
 research, artificial intelligence for, 1773
 types, 1768
 Forensic pathology, 1768, 1772, 1778
 Forensic physician, 1783
 Forensic reasoning, 1771
 Fourier transform algorithm, 439
 Foveated kernel, 93
 Fractal dimension, 787
 Fracture detection, 877
 Fragment assembly, 664
 Framingham Heart Study, 817
 Frequent association rule mining, 305–306
 Frontiers in AI, 124–125
 Frontline Health Worker Virtual Health Assistance, 619
 Frontotemporal dementia (FTD), 1678
 Full omics analysis, 724
 Fully convolutional network (FCN), 543
 Fully convolutional neural networks, 1509
 Functional impairment, 1632
 Functional MRI (fMRI), 1586
 Functional thyroid disorders, AI application, 681–682
 Fundus photograph, 1063, 1554, 1555
 diabetic retinopathy, 1555
 glaucoma, 1557, 1558
 Future Emerging Technologies (FET), 1018
 Future of AI, 125–126

- Fuzzy clustering, 1496
Fuzzy logic, 164, 1794
Fuzzy-logic systems, 186–187
Fuzzy sets, 164
Fuzzy systems, 1447–1450
- G**
4G, 1231
5G, 1231, 1232
Gabor functions, 788
Gabor kernels, 82
Gait rehabilitation, 1813–1814
GAMEPHARM platform, 1799
Gamification, 1815
Gas sensor array, 1204
Gas sensors, 1204, 1206, 1210, 1214
Gastric metaplasia (GM), 958
Gastroenterology, 1239
Gastrointestinal (GI), 924–927, 930, 959
 automatic report generation, 929
 CAD systems, 922, 930
 classification and segmentation, 931
 clinical verification/emerging commercial systems, 931, 932
 deep computer vision-based approaches, 933
 deep learning-based approaches, 923, 924
 endoscopy, 921, 924–925
 generalizability, 928
 hand-crafted-feature-based approaches, 923
 metrics/evaluation, 924–928
 segmentation, 926
Gated Recurrent Units (GRUs), 586
Gates, Bill, 294
Gaussian mixture model (GMM), 776, 1433
Gaussian potential function (GPF), 282
Gaussian process regression (GPR), 1281, 1282
Gendercide, 1022
Gender identity, 399
Gene editing, 298
Gene editing technologies, 298
General AI, 7
General Data Protection Regulation (GDPR), 141, 353, 354, 1475, 1822
Generalizability, 506
General practice, 713, 714, 717, 734
Generative adversarial network (GAN), 667–669, 1542
Generative adversarial Networks (GAN), 90, 924, 956, 979, 1130, 1516
Generative network, 1131
Generative pre-trained transformer (GPT-3), 19, 476
Generative Text Compression with Agglomerative Clustering Summarization (GTCACS), 1126, 1130, 1140
Genetic Algorithm-based Search (GA), 790
Genetic algorithms, 70
Genetic editing technology, 1023
Genetic engineering (GE), 295, 297, 1021
Genetics & molecular markers, 989
Genetic sequencing, 723
Genome wide association studies (GWAS), 1075, 1414, 1416
Genomic medicine
 AI interpretation, 1088, 1094
 cancer, 1088
 coding region, variants, 1088, 1089
 definition, 1088
 diagnosis (phenotyping), 1091
 human genome, 1087
 interpretation of variants, NLP, 1090
 non-synonymous variants, 1089, 1090
 optimal drug treatment, 1092, 1093
Genomics, 723, 1062, 1236
Genomics of Drug Sensitivity in Cancer (GDSC), 1092
Genomic studies, 348
Genotypic resistance patterns, 1332
Gestational diabetes, 676, 703
Glaucoma, 1556–1560
 anterior-segment OCT, DL algorithms, 1528–1529
 DL algorithms, optic dis-centred OCT, 1538–1539
 electronoc health records, 1540
 fundus photographs, 1520–1524, 1528–1529
 visual fields, 1539
Gleason Score, 864
Glioblastoma multiforme, 529
Glioma brain tumous, 529
Global Distance Test (GDT) metric, 666
Global health, 619
Glomerular filtration rate (GFR), 562
Glomerulus, 541–546
Glucose metabolism (GM), 1585
GLY1 and GLY2 nano-lipid carriers, 1179
Good old-fashioned AI (GOFAI), 1597
Google, Amazon, Facebook and other Apple (GAFAMS), 1350
Google Cloud AutoML Vision (Google), 475
Google Flu Trends, 1345
Google Inception v3, 1266
Gout, 774, 779
Governance, 616–618
Gradient boosted machine (GBM), 1402
Gradient boosting, 1333
Graft versus host disease (GVHD), 1404
Graph convolutional networks (GCN), 249, 1093
Graph convolution neural network (graph CNN), 1076, 1391
Graphical processing units (GPUs), 6, 475, 761, 1709
Graph neural network, 275
Graves' disease, 681
Gray-level co-occurrence matrix (GLCM), 958
Gray Level Co-Occurrence Matrix (GLCM), 790, 1324
Ground truth, 865
Group B streptococcus (GBS), 1036
Group level fairness, 402
Grover's search algorithm, 439
GUIDON, 334
Guillain-Barre syndrome, 1763
Gut microbiome, 723

H

Hackathon, 364
 agenda, 367
 effectiveness and usefulness, 367
 format, 366
 goal and theme, 365
 ideation, 366
 involvement, 368
 leverage existing resources, 368
 limitations, 368
 mentoring, 367
 participants, 366
 problem-solving, 364
 resources, 366
 significance, 365
 stakeholder involvement, 366
 team formation, 366

Hadamard gate, 433, 434

Haem-oncology screening, 1430

Haemo-parasitology, 1434

Haemorrhagic stroke, 1504

Hallucination, 1596, 1598, 1604

Hanover project, 64

Hard parameter sharing, 414

Harmonization, 1474

Hausdorff Distance, 1515

Head & Neck cancer

- auto-segmentation, 988
- definition, 987
- histopathology, 988
- multippectral imaging, 989
- prognostication, 987
- radiological staging, 988
- radiomics, 987
- treatment response, 988
- treatment toxicity, 988

Health blogs

- adverse drug reaction, 1127–1128
- compared systems, 1132–1133
- data analytics, 1136–1140
- data filtering, 1129–1130
- data gathering, 1128–1129
- dataset and diseases, 1132
- evaluation strategy, 1133–1134
- social analytics for healthcare, 1127
- topic analysis and sub-clustering, 1134–1136
- topic modeling, 1130–1132

Healthcare, 6, 12, 707, 760, 764, 767, 1277, 1283, 1610, 1611

- data, 1399
- professionals, 705, 1337
- service redesign, 124
- system, 703

Healthcare-associated infections (HCAI), 1332, 1336

Healthcare ethics and artificial intelligence

- beneficence, 140–141
- justice, 142
- nonmaleficence, 141–142
- respect for autonomy, 139–140

Health insurance, 598, 1613

Health Insurance Portability and Accountability Act (HIPAA), 146, 1822

Health Insurance Portability and Accountability Act of 1996 (HIPAA), 353

Health Insurance Portability and Accountability Act Privacy Rule (HIPPA), 141

Health recommender systems (HRSS), 276

Health record synchronization, 724

Health tools and applications, 1346

Healthy Controls (HCs), 1581

Healthy Moms program, 1614

Hearing loss, 1706, 1707

Heart failure, 1480–1484

- cardiac dysfunction, 818

- common causes, 817

- diagnosing, 817

- morbidity, 817

- prognosis, 817, 818

Heart failure with mid-range ejection fraction (HFmrEF), 817

Heart failure with preserved ejection fraction (HFpEF), 817

Heart failure with reduced ejection fraction (EFrEF), 817

Heart rate, 1145

Heart rate characteristics (HRC), 1050

Heart rate variability, 1146, 1149, 1151

Heart rhythm, 1146

Heat syndrome, 1256

Heisenberg's uncertainty principle, 431

Helicobacter pylori (HP), 943

Helper T Lymphocytes (HTL) epitopes, 1390

Hematocochlear test, 809

Hemodialysis (HD), 580, 1442, 1444, 1445, 1449, 1450

Hemoglobin (Hb), 581

Hemoglobin prediction, 581

- MPC, 581

- Random Forest, 586

Hemorrhagic stroke, 1734

HemoScreen, 1436

Hepataug, 946

Herb-Target Interaction Network (HTINet) approach, 1255

Heterogeneous datasets, 1337

Heterogeneous treatment effects (HTE), 690, 693, 694

Hierarchical agglomerative clustering (HAC), 1078

- qglomerative clustering, 1080

- hierarchical clustering algorithm, 1079

Hierarchical clustering, 1496

Hierarchical feature extraction, 41, 49

High-performance algorithm, 697

High-performance computing, 1276, 1278, 1280, 1281

High-resolution peripheral quantitative computer tomography (HRpQCT), 781

High throughput screening (HTS), 1491

Hilbert's space, 433

Hip and pelvis, 496

Hip angle measurements, 497

HIPPO module, 496

Histogenomics, 1268

- Histogram of oriented gradients (HOG), 541, 1794
Histogram semi-Markov models (HSMM), 1794
Histology, 540
Histopathology, 763
HIV, 1332, 1335
HLA typing, 1389, 1390
Homomorphic encryption (HE), 152, 153
Honey, P., 322
Hopfield-like network ('HLN'), 26
Hormone, 674, 675, 681, 682
Hospital-driven data, 1391
Hospital setting, 1335
Host-pathogen interactions, 1370
HSC lineage engineering, 1101
HUGO project, 60
Human-Centered Artificial Intelligence (HAI), 1390
Human decision making, 132
Human immunodeficiency virus (HIV-1), 1434
Human interpretation procedures, 1471
Human intestinal absorption (HIA), 638
Human learning, 408
Human Leukemia Antigens (HLAs), 1388, 1390
 Class I, 1391
 Class II, 1391
 methods, 1391
 rare alleles, 1391
 variation, 1391
Humanoid nurse robots (HNRs), 753–755
Human Phenotype Ontology (HPO), 1402
Human protein-protein interaction network, 1081
 interactome, 1074, 1075, 1078, 1083
Human reasoning, 375
Human-robot interaction, 1814
Human trial, 1393
Human visual perception, 90
 color processing and colorization, 90
 foveated vision, 91
Humoral immune response, 1388
Hurufism, 205
Hydrophobic effect, 662
Hyperspectral Imaging (HSI), 848
Hypertension, 735
 AI approaches, 690
 BP measurements, 690
 management, 690
 paradox, 690
 prediction models, 695
HyperTrak scanner, 103
Hypoglycemia, 703, 707, 708
Hypotension prediction index (HPI), 1459
- I**
iCub robot, 1600
IDEAS® software, 1433
Identification, 909
Idiopathic arthritis, 777, 780
IgA nephropathy (IgAN), 545
Image analysis, 523, 530, 1721
Image-based automatic classification of malaria, 1357–1361
 data classification, 1357
 data processing, 1357
 mobile applications and end-to-end systems, 1360–1361
 Plasmodium falciparum detection, 1359–1360
 Plasmodium life stages, 1360
Image-based automatic classification of mosquito vectors
 mosquito behavioral pattern, 1363–1364
 mosquito species identification, 1361–1363
Image-based diagnosis, 1063
ImageNet, 1599
ImageNet Large Scale Visual Recognition Challenge (ILSVRC), 1710
Image processing analysis, 974
Image quality assessment, 1540–1542
Image quality tools, 528
Image recognition, 556, 722
Image segmentation, 827, 1279
Imaging data, 906
Immune checkpoint blockade, 1430
Immune epitope database (IEDB), 1391
Immune globulin, 1402
Immunoinformatics, 1389
Immunological Elastic-Net (iEN), 1405
Immunology, 1388, 1390
Impaired Glucose Tolerance (IGT), 276
Implantable cardioverter-defibrillator (ICD), 1480–1482, 1484
Implantable computing systems, 1194
 autonomous pacemakers, 1194
 implantable cardioverter defibrillators, 1194
 safety testing, 1198
Improved supervised normalized cuts (ISNC)
 segmentation, 1102
Inattention monitoring, 1802
Independent component analysis, 1496
Individualization, 595
Individual level fairness, 402
Individual treatment, 728
Inductive bias, 194
Industrial Revolution (IR), 744–745
Inertial measurement units (IMUs), 842
Infantile facial video recording, 1540, 1541
Infection biology, 1370–1371
 macroscale and digital biomarkers, 1373–1374
 mesoscale and aspects of temporal dimensions, 1373
 molecular, 1374–1375
 nano-and microscale, 1371–1373
Infection phenotypes, 1373
Infection prevention and control (IPC), 1336
Infectious disease(s), 345–346, 1370, 1373–1375
 AI tools, 1329, 1333
 causative organisms characteristics, 1328
 clinical diagnosis and management, 1333
 decision-making, 1329, 1331, 1338
 diagnosis, 1333, 1334
 diagnostics, 1374

- Infectious disease(s) (*cont.*)
 drug design, 347
 drug resistance, 346–347
 dynamic variables, 1328
 emerging pathogens, 347–348
 interactions, 1330
in-vitro, 1329
in-vivo, 1329
 management, 1328–1330
 non-physician healthcare professionals, 1331
 patient factors, 1328
 primary care physicians, 1329
- Superior alveolar nerve (IAN), 902
- Infertility, 868, 869
 and AI, 1008, 1011
- Inflammatory bowel disease (IBD), 1404
- Inflammatory bowel disease unclassified (IBDU), 1040
- Information theory, 1693, 1697
- Informative frame selection, 830
- Informed consent, 140
- Innate immune system, 1388
- Innovation, 115, 364–366, 368
- Inotropes infusion, 1458
- In silico* approaches, 1491
- Institute of Medicine (IOM), 1400
- Insulin, 706
- Insurance systems, 442
- Integrative approach, 1418
- Integrative Sparse Classification*, 798
- Intellectual property (IP), 368
- Intelligence quotient (IQ), 1582
- Intelligent machines, 1755
- Intensity modulated beam techniques, 1282
- Intensive Care Unit (ICU), 1400, 1470
- Interactome hierarchy (I-T) labeling, 1081
 unexpected neighborhoods, 1083
- Interactome Taxonomy (I-T), 1077, 1078
- Inter-annotator agreement (IAA), 273
- Interaural difference (IAD), 1708
- International Collaboration for the Automation of Systematic Reviews (ICASR), 258
- International Committee Monitoring ART (ICMART), 1009
- International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10), 1597
- International Telecommunication Union, 1233
- Internet of things (IoT), 130, 1232, 1233, 1276, 1280
 based monitoring, 676
- Interoperability, 463
 standards, 1344
- Interpretability, 1471, 1476
- Interstitial lung disease (ILD), 762
- Interventional radiology (IR), 460
 challenges, 464–465
 decision support, 466–467
 image acquisition and processing, 469
 IR suite, 466
 non-vascular interventions, 465
- opportunities, 465
 patient monitoring and procedural support, 468–469
 periprocedural support, 468
 prevention of errors, 467–468
 prognostication and outcome prediction, 469
 residency and fellowship training, 469–470
 triaging of patients, 467
 vascular interventions, 465
- Interventional Radiology suite (IR suite), 466
- Intestinal metaplasia (IM), 958
- Intractable diseases, 1088
- Intraoperative hypotension, 655
- Intra-operative image analysis, 830
- Intra-operative registration, 828
- Intraoperative video analysis, 859–860
- Intravenous alteplase administration, 1734, 1737
- In-vitro fertilization (IVF)
 applicability of AI for, 1004–1005
 cycles, 1008, 1011
 ethical challenges, 1005
- In-vitro* fluorescence activated cell sorting (FACS) process, 1103
- In vitro–in vivo extrapolation (IVIVE), 1498
- Iron deficiency anaemia (IDA), 1429
- Ischemic heart disease, 814–816
- Ischemic stroke, 1734
- Iterative dichotomizer 3 algorithm, 12
- J**
- Jaccard Index (JI), 974, 1515
- Java, 8
- Jaw, 909, 912, 914
- Justice, 142
- K**
- Kalman filtering, 1559
- Kawasaki's disease (KD), 1402
- Kellgren Lawrence (KL) grading system, 781
- Kellgren-Lawrence score, 495
- Kenya Medical Supplies Authority (KEMSA), 620
- Keratocystic odontogenic tumors (KCOTs), 912
- Kernel function, 195
- Kidney cancer, 866
- Kidney pathology, artificial intelligence
 classification and identification of specific components, 545
 classification based on pathological category, 545
 classification of images with immunohistochemistry, 545–546
 classification of major pathological findings, 544
 clinical category and genotype, classification based on, 546
- conventional features, 541
- deep learning, 542
- detection, 541–542
- segmentation of glomeruli, 542–543
- segmentation of multiple structures, 543

- Kinect system, 1792, 1793
Kinetic properties, 1174
K-Nearest Neighbor (KNN), 14, 1495
algorithm, 815
classification, 1332
Knee arthroplasty, in orthopaedics, 879
Knee osteoarthritis, 495
Knowledge base (KB)
ontological, 163
rules-based, 163
Knowledge-based diagnosis, 185
fuzzy-logic systems, 186–187
ontology based systems, 187–188
rule-based diagnosis, 186
Knowledge-based systems, 1335
Knowledge engineering (KE), 1250
Knowledge expansion, 277
external knowledge, 279, 280
mining potential, 280
Knowledge graph, 269
Knowledge mining, 268, 272, 277, 280
Knowledge representation and reasoning, 163
Kolb's model, 322
Kronecker product, 436
- L**
Labelling, 907
Labels, 1568, 1570
Laboratory information systems (LIS), 806
Laboratory medicine, 804
machine learning implementation in, 805–806
machine learning models in, 806–810
LAMA, 497
Landsat TM data, 631
Large vessel occlusion (LVO), in prehospital stage, 1737
Large vessel occlusion (LVO) detection, 1743
Laryngology
audio based prediction, 995
clinical based prediction, 995
image based prediction, 995
video based prediction, 995
Video fluoroscopic swallow (VFS), 995
voice/larynx, 995
Laser Desorption Ionization Mass Spectrometry (LDI-MS), 846
Lasso regression, 1431
Latent Dirichlet Allocation (LDA), 1132
Law's Texture Energy Masks, 790
Layer-wise relevance propagation (LRP), 507
Layer-wise relevance propagation (LRP) method, 291
L-Dopa-induced dyskinesias, 1685
Leabra architecture, 25
Learned iterative shrinkage and thresholding algorithm (LISTA), 46–48
Learning from demonstration (LfD), 1813
Learning healthcare systems (LHS), 597, 1400–1401
Learning health systems, 62–63
Learning to Learn, *see* Meta learning
Least absolute shrinkage and selection operator (LASSO), 1117, 1335
Left ventricular ejection fraction (LVEF), 817
Legal liability of AI, 608
Leucocytes, 1429
Leukemia, 1110, 1431
Lexical analysis, 1347
Liberal eugenics, 296–298
Lifestyle-related diseases, 1088
Ligand screening and pharmacology
ligand-based approach, 637
structure-based approach, 637
Likelihood mining criterion, 307–308
Limits, 376
Linear discriminant analysis (LDA), 848
Linear regression (LR), 12, 583
Liquid-based cytology (LBC), 1319
Lisp, 8
Literacy, 116–118
Liver disease comorbid network (LDCN), 1252
Liver-yang syndrome, 1255
Living Systematic Reviews, 258, 265
Local binary patterns (LBP), 541, 788
Local interpretable model-agnostic explanation (LIME) algorithm, 291, 507, 694
Localization, 923
Locally Interpretable Model-agnostic Explanations (LIME), 930
Locally Weighted Scatterplot Smoothing (LOWESS) algorithm, 246
Logic-based methods, 162
knowledge representation and reasoning, 163
language of logic, 162–163
Logistic model tree machine learning (ML) algorithm, 1529
Logistic regression, 12
Log likelihood approximation residual-based decision function (BBLL-R), 795
Log likelihood sparsity-based decision function (BBLL-S), 796
Longevity medicine, 1158
AI applications in, 1165–1166
future research, 1166
physicians, 1161–1164
and public health, 1164–1165
Longitudinal dataset, 1693, 1695
Long short-term memory (LSTM), 247, 586, 692, 779, 1375, 1380
Long short term memory (LSTM)-based meta learner network, 412
Long short-term memory (LSTM) network model, 844, 1061
Lou Gehrig's disease, 1763
Lower extremity leg, 498
Lower Respiratory Tract Infections (LRTI), 1331

Low-middle income (LMIC), 1336

Low-resource, 616

Lumiata, 64

LUR methods, 627

Lymphoma, 1431

Lymphoproliferative disorders, 1432

M

Machine learning (ML), 4, 39–41, 51, 108, 116, 121, 122, 124, 164, 165, 210, 212, 218, 240, 259, 263, 294, 298, 299, 356, 409, 460, 461, 488, 504–505, 512, 523, 528, 530, 532, 540, 554, 557, 674, 676–678, 680–682, 690, 698, 702, 747, 760, 774–776, 779–781, 805, 856–858, 860, 864, 866, 870, 890, 891, 899, 901–903, 906, 920, 942, 984, 1030, 1100–1101, 1204, 1205, 1211–1213, 1221, 1224, 1225, 1230, 1231, 1233, 1236, 1238–1242, 1248, 1250, 1265, 1343, 1347, 1370, 1389, 1391, 1392, 1398–1404, 1454–1456, 1459–1461, 1470, 1491, 1492, 1500, 1529, 1534, 1539, 1559, 1581, 1596, 1598, 1599, 1601–1605, 1622, 1664, 1681, 1692, 1745, 1756, 1759–1761, 1810, 1812–1816

advantages, 1283

adverse events, 219

algorithms, 1158, 1160, 1281

algorithms on neurosurgery, 1664

AKI onset prediction (*see* Acute kidney injury (AKI) prediction)

ALS study with, 1694–1697

AM data analysis, 1251

ANNs, 1494–1495

Ayurvedic medical system, 1252

based imaging tools, 1734

basic concepts, 1277

beam angle optimization, 1279

case example, 9–10

cell segmentation, 1669

challenges, 1625–1626

chemical toxicity prediction, 1493

CLABSI, 219

classifiers, 1669

collecting balanced datasets, 401

continuous monitoring of, 403–404

cycle, 1427

data and model documentation, 402

data mining, 1251

data sets, 1623

definition, 984

in depression and anxiety, 1568–1576

diagnostics, 219

disaggregated data, 402

for drug dependence, 1623–1625

EHR, 218

external auditing framework, 403

fairness aware models, 402–403

illustration of, 9

implementation in laboratory medicine, 805–806

issues, 401

Kmeans clustering algorithm, 1668

limitations of, 16

medical errors/polypharmacy, 220

methods, 1076, 1077

models in laboratory medicine, 806–810

in nanomedicine, 1175–1179

for nanorobotic surgery, 1181–1183

for nanotoxicology, 1180

NLP, 219

in orthopaedics, 875

in precision nanotheranostics for cancer, 1179–1180

in predictive analytics, 881–883

problem solutions, 11

process, 9

and quantum enabled technologies, 1181

and regenerative nanobiology, 1181

reinforcement, 985

reinforcement learning, 15–16

supervised, 985

supervised learning, 165, 1493–1494

supervised learning algorithms, 11–15

SVMs, 166

template, 985

test dataset, 1491, 1492

traditional medicine, 1251

training dataset, 1491

types of, 10–11

unsupervised, 985

unsupervised learning, 166, 1494–1496

unsupervised learning algorithm, 15

workflow, 1281

Machine learning, medical diagnosis

deep learning approaches, 194

formalisms, 195–196

as function approximation, 193–194

importance of data, 196

learning from data, 191–193

SVM, 194, 195

Machine learning (ML) algorithms, 1412

challenges and promising research, 1418–1420

clinical management, 1416

feature extraction, 1417

reinforcement learning, 1413

semi-supervised machine learning, 1413

supervised learning, 1413

unsupervised learning, 1413

Macro-level barriers, 1336, 1337

Magnetic resonance imaging (MRI), 469, 504,

827, 1032, 1265, 1296, 1304, 1373, 1572,

1575, 1581

automated analysis, 1718, 1719

challenges, 1718, 1720

detection and segmentation, 1721

modalities, 1718, 1719

neurological cancers, 1718

T1, 1718

T2/FLAIR, 1718

Major adverse cardiac events (MACE), 858

Major Histocompatibility Complex (MHC), 1388, 1390

- Malaria, 1354
 applications of artificial intelligence, 1357–1366
 deployment of artificial intelligence, 1356–1357
 diagnosis, 1354–1355
 image-based automatic classification (*see* Image-based automatic classification of malaria)
- Mammographic Image Analysis Society (MIAS)
 dataset, 1289
- Mammography, 1303
- Mandibular canal (MC), 909
- Markers of disease, 1210
- Markov Decision Process (MDP), 410, 1334
- Markov model, 696
- Markov random field (MRF), 281
- Massachusetts General Hospital, 1692
- Massive data, 1343
- Mass spectrometry (MS), 242, 1331
- Maternal health, AI and machine learning for, 1008
- Mathematical oncology, 343–345
 precision medicine, 344–345
 prediction of cancer risk, 344
- Matrix-assisted laser desorption ionization-time of flight (MALDI-TOF) mass spectrometry, 1331, 1356, 1364, 1365
- Matthews correlation coefficient (MCC), 921
- Maxillofacial surgery, 888
 artificial neural networks, 893
 challenges, 898–899
 computer vision, 894
 facial paralysis gradation, 902
 impacted teeth and minor oral surgery, 902
 implant surgery, 900
 machine learning and deep learning, 891
 maxillofacial pre-surgical imaging, 899
 natural language processing, 893–894
 oncosurgery and reconstruction, 901
 orthognathic surgery, 899–900
 pain control, 902
 role of surgeons, 897–898
 SCT, 903
 surgeons dilemmas, 895–897
 surgical data science, 902
 surgical scene analytics, 903
 TMJ surgery, 900–901
 trauma surgery, 901–902
- Maxillofacial traumatic injuries, 901
- Maximally Stable Extremal Region (MSER)
 algorithm, 1321
- Maximum a Posteriori decision function (BBMAP), 795
- Mayo Clinic Anemia Management System (MCAMS), 582
- mDiagnostics, 1240
- Mean absolute error (MAE), 627, 692
- Measurements of Pollution in the Troposphere (MOPITT), 630
- Mechanical thrombectomy, 1743
- Mechanical ventilation, 1473
- Media attention, 1337
- Median filtering (MF), 1211
- Medical AI
 business perspective, 604–605
 co-creation, 607
 consumer perspective, 605–606
 ethics, law and policy, 608–609
 multi-stakeholder engagement, 607
 myth of generalizability, 606–607
 value metrics, 607
- Medical applications, 273
 dermatosis diagnosis, 276
 disease diagnosis, 274
 disease management, 276
 disease risk prediction, 275, 276
 EMR quality control, 275
 over-testing detection, 275
- Medical artificial intelligence (AI)
 anonymization, 148
 applications, 147
 dataset type, 148
 FL, 149
 guarantees, 147
 privacy, 147
 research, 146
 technical standards, 147
 trustworthy, 147
 verifiability, 147
- Medical data, 1717, 1727
- Medical datasets, 306
- Medical decision-making, 232
- Medical decisions
 Bayesian models for decision support, 169–171
 causality, 171–173
 combinatorial optimization methods, 168–169
 deep learning revolution, 167–168
 explainability, interpretability and fairness, 173–174
 logic-based methods, 162–164
 statistical modelling, 164
 taxonomy, 161–162
 three machine-learning approaches, 165–167
- Medical decision support systems (MDSS), 1472
- Medical diagnosis
 clinical data application, 247
 diagnostic reasoning, 182–185
 knowledge-based diagnosis, 185–188
 machine learning for diagnosis, 191–196
 model based diagnosis, 188–192
 omics data application, 247, 248
- Medical education, 322, 328, 336
- Medical epistemology and AI
 computational biases, 378–379
 data, 373–374
 data curation and use, 375–376
 data-utopianism, 374–375
 human biases and prejudices, 377–378
 limits, risks and biases, 376–379
- Medical ethics and AI, 379–380
 digital capitalism, 381–382
 patient-doctor relationship, 380–381
- Medical history, 725

- Medical imaging, 220, 460–463, 760, 761, 766, 780–781
 AI-enabled technology towards, 503–504
 in clinical diagnosis, 502
 explainable artificial intelligence, 505–506
 machine learning, 504–505
 post-hoc interpretability, 507
 transparent models, 506
- Medical informatics, 733
 AI models, 250
 clinical data, 241
 data processing, 244
 EHRs, 240
 explainability, 249
 patient-related information, 240
- Medical Information Mart for Intensive Care (MIMIC), 1334
- Medical knowledge graph (MKG), 277
- Medical knowledge management, 68
- Medical knowledge network (MKN) model, 281
- Medical knowledge representation, 270–272, 284
- Medical ontology, 271–273, 284
- Medical professional liability (MPL), 222
- Medical records, 725
- Medical robotics, 826
- Medicinal chemist, 1392
- Medicine
 AI, 32, 207
 ANNs, 22
 automation, 209
 automatons, 204
 CCA1, 26–31
 cognitive architecture, 24
 cognitive architectures, 31
 computer, 210, 211
 computerized medical diagnosis, 209
 conceptual revolutions, 208
 deep learning, 31, 32
 DENDRAL expert system, 22
 DNN, 23
 examples, AI, 210
 myth, 206
 perceptron, 23
 personal computer, 208
- Medicine, AI
 deep learning, 456
 DNN, 450, 455
 genomics/epigenomics, 455
 machine learning techniques, 449
 medical imaging, 455
 natural language processing, 455
- MedicMind Deep Learning Training Platform (MedicMind), 475
- Medico-administrative data, 1343
- Ménière's disease (MD), 1706
 artificial intelligence, 1710–1715
 diagnostic method and dilemma, 1708
 history of inner ear MRI, 1708
 medical image analysis, 1709–1710
 prevalence, 1706
 sequence and analysis of inner ear MRI, 1708–1711
- Mental health, 1633, 1638
- Mental Health Gap Action Programme (mhGAP), 1615
- Mentoring, 367
- messenger RNA (mRNA), 1113
- Meta-analysis, 326, 328–336
- Metaheuristics, 70
- Meta learner model, 409
- Meta learning, 408, 409
- Metal oxide semiconductor (MOS) sensors, 1210
- Meta reinforcement learning, 410
- Meta-training dataset, 409
- Meteor Nexus, 639
- Methicillin resistant *Staphylococcus aureus* strains (MRSA), 1332
- Metric-based approach, in deep learning, 411
- mHealth, 1230, 1233, 1243
 cardiovascular medicine, 1236–1237
 consensus evaluation, 1241–1242
 dermatology, 1240
 endocrinology, 1239–1240
 evidence-based medicine, 1235
 gastroenterology, 1239
 genomics, 1236
 infectious diseases, 1242
 neuroscience and neuropsychiatric disorders, 1238
 obstetrics and gynaecology and pediatrics, 1241
 respiratory medicine, 1237–1238
 rheumatology, 1238–1239
 services, 1233
 urology, 1239
- Microarray-based transcriptome technologies, 1112
- Microbes, 1370
- Microbiology, 1329, 1331–1333, 1336–1338, 1373
- Microbiome analysis, 1404–1405
- Microdissection testicular sperm extraction (micro-TESE), 869
- Micro-level barriers, 1336, 1337
- Microns per pixel (MPP), 530
- Micro-organisms, 1331
- Microscopic parasites, 1370
- Microscopy, 1370
- Middle East Respiratory Syndrome (MERS), 1347
- Mie theory-based blood analyser, 1429
- miLabTM platform, 1361
- Millennium Development Goals, 1234
- Minimally invasive surgery (MIS), 840, 843, 942, 1269
- Minimum operable quantity (MOQ), 276
- Minimum viable product (MVP), 607
- Ministry of Health (MOH), 217
- Minor oral surgery, 902
- Min-support, 304–308
- Mismatch-based reperfusion therapy, 1741
- MIT Hacking Medicine, 364
- ML/AI data analytics, 1471
- ML-driven models, 1414, 1415
- ML prediction model, 694
- MNIST dataset, 77
- Mobile applications for malaria diagnosis, 1360
- Mobile health (mHealth), 1221
- Mobile Stroke Units (MSUs), 1512

- Model-agnostic meta learning algorithm (MAML), 412
Model-based algorithms, 184
Model-based approach, in deep learning, 412
Model based diagnosis, 188
 abductive diagnosis, 188–189
 Bayesian diagnosis, 189–191
 causal reasoning for diagnosis, 191–192
Model driven-AI, 290
Model explainability, 332, 1764
Model predictive control (MPC), 581, 1446, 1448
Modern medicine, 5
Modified Rankin Scale (mRS), 1511
Molecular biology, 1373
Molecular data, 1392
Molecular infection biology, 1374–1375
Molecules, 636, 639
Monitoring/assessing surgical procedures, 830
Monte Carlo simulation, 1279
Moorfields DL system, 1562
MOPITT sensor, 630
mOverallyHeart, 1237
Moving window average (MVA), 1211
MPC-based approaches, 581
mPlatform, 1242
Multi-criteria decision-making (MCDM), 1403
Multi-disciplinary, 364, 366
Multi-drug resistant (MDR), 345, 346
Multi-factorial threshold model, 1581
Multilayer perception (MLP) regression, 17–18, 626, 868, 1445
Multi-layer perceptrons (MLP), 41, 583
Multi-leaf collimator, 1282
Multimodal method, 638
Multi-omics late integration (MOLI) method, 1092
Multi-parameter immune-phenotyping model, 1103
Multiple adaptive linear elements (MADALINE), 1426
Multiple kernel learning (MKL), 1068
Multiple myeloma, 1432
Multiple sequence alignment (MSA), 664, 1089
Multi-pooling operation, 277
Multi-scale *in-silico* computational model, 1099
Multi-scale Temporal Memory (MTM), 247
Multispectral or hyperspectral imaging, 989
Multi-stakeholder engagement, 607
Multi-task learning, 413–416, 637
Multi-Task Network Cascades, 414
Multivariable models, 1345
Multivariate inference methodology, 282
Mumford, A., 322
Musculoskeletal diseases, 490–491
 AI-driven KOALA software, 495
 AI in computed radiographs, 493
 anomaly detection, 495
 building and validating AI models for radiographs, 493–494
 CR images, 492
 CT and MRI imaging, 492
 disease management into digital age, 491
 HIPPO module, 496
 LAMA, 497
 non-structured and manual workflows, 498
 PANDA, 495
 quality control in imaging, 494–495
 radiology, 491–492
Musk, Elon, 294
Mutation, 1390
Mycobacterium tuberculosis, 1332
MyCOPD, 1237
myDiabetes, 1240
Myocardial infarction (MI), 815, 1480–1483
- N**
- Naïve Bayes classifier, 13–14, 1495
Naïve Bayesian nearest neighbour algorithms, 1793
Nano-biointerfaces, 1175
Nanocarrier drug delivery, 1178
Nanodrugs, 1178, 1179, 1181
Nano-ethical perspective, of nanomedicines, 1183
Nanomedicine, 1170
 machine learning in, 1175–1179
 nano-ethical perspective, 1183
Nano neurosurgery, 1182
Nanorobotic surgery, 438, 441
 machine learning for, 1181–1183
Nano-robotic variations, 1182
Nanotechnology, 1170
Narrow AI, 7
Narrow-band imaging (NBI), 942, 957
National Cancer Institute's Cancer Genome Atlas project, 1236
National Centre for Complementary and Integrative Health (NCCIH), 1248
National Health Service (NHS), 355, 357
National Lung Cancer Screening Trial, 762
National Science Foundation Network (NSFNET), 210
Natural language processing (NLP), 19, 116–119, 141, 218, 262, 294, 333, 464, 571, 655, 779, 858, 893–894, 984, 1265, 1400, 1403
Natural reference methods, 1344
K-nearest-neighbors (KNN), 1118
NEAT method, 416
Neocytolysis, 582
Neonatal intensive care unit (NICU), 1030
Neoplastic bone marrow pathology, 1432
Nephropathology, 540–543, 546
NetMHCIPan, 1391
NetMHCPan, 1391
Network Medicine, 1074
 machine learning challenges, 1075
Network pharmacology, 1255
Network-proximity principle, 1075, 1077, 1082
Neural architecture search, 416–418, 477
Neuralink platform, 435
Neural nets, 102
Neural network-based tools, 1390
Neural networks, 38, 121, 167, 416, 489, 856, 902, 1598, 1599, 1602–1604
 architecture of, 41–44
 generic iterative algorithm, 48–49

- Neural networks (*cont.*)
 geometric understanding, 51–52
 LISTA, 46–48
 representation power of deep, 44–45
- Neuroimaging data, 1634–1636
- Neuroimmunological diseases, 1110
- Neurological Clinical Research Institute, 1692
- Neurological treatment, AI
 accurate detection in prognosis, 1670
 AI-driven solution, 1670–1671
 challenges of clinical observation, 1670
 childhood medulloblastoma, 1665–1667
 data collection, 1667
 data pre-processing, 1667
 dividing training and test data, 1669
 feature extraction, 1668
 feature reduction, 1668
 feature selection, 1668
 ground truth annotations, 1668
 methods and materials for CNS tumors, 1665–1670
- Neurophysiology, 726, 1756
- Neuropsychology, 1811–1812
- NeuroQWERTY index, 1682
- Next generation machine learning, for nanorobotic surgery, 1181–1183
- Next generation sequencing (NGS), 242, 780
- nHLAPred, 1391
- Niels Bohr model, 425
- NIROM, 628
- Nitric oxide (NO), 765
- node2vec, 1076
- Non-blood flow features, 1150
- Non-communicable diseases (NCDs), 713
- Non-compartmental analysis (NCA), 1490
- Non-contrast CT (NCCT), stroke assessment, 1506
- Non-generative network, 1131
- Non-Hodgkin's lymphoma, 1431
- Noninvasive biopsy methods, 1012
- Non-invasive monitoring, 1483
- Nonmaleficence, 141–142
- Non-negative matrix factorization (NMF), 776, 1132
- Non-obstructive azoospermia (NOA), 868
- Normalization, 246
- Normalized cross correlation (NCC), 848
- Normalized difference built-up index (NDBI), 632
- Normalized Pointwise Mutual Information (NPMI), 1136
- N-Shot Learning, *see* Few Shot Learning
- Nurse-robot interface, 754
- Nursing, 752
 education, 753
 practice, 752–753
- Nutrition management, 276
- O**
- Objective Structured Assessment of Technical Skills (OSATS), 334, 831
- Objective Structured Clinical Examination, 1241
- Objectivity, 373
- Observational data analysis, 1344
- Observational Medical Outcomes Partnership (OMOP), 1344
- Obstetrics and gynecology, 1004
- Occlusion test, 1555
- Occupational rehabilitation, 1812
- Old school vaccine, 1388
- OMERACT-EULAR Synovitis Scoring (OESS), 781
- Omics data, 640
- Omics Data Management (ODM), 241
- Oncology, 736–737
- Oncosurgery and reconstruction, 901
- One-class logistic regression, 1101
- One-hot encoding, 1514, 1515
- On Generation of Animals*, 1018
- Ontology and knowledge graph, 1250
- Ontology based systems, 187–188
- Oocyte inspection, 1004
- Opaque model, 507
- OpenPose platform, 1791, 1794
- Open science approach, 1823
- Open source software (OSS), 1476
- Ophthalmic surgery, 836
- Opsonins, 1178
- Optical chromoscopy techniques, 952
- Optical coherence tomography (OCT), 211, 958, 992, 1535, 1555
 age-related macular degeneration, 1561
 DL algorithms for retinal diseases, 1535–1538
 glaucoma, 1528–1529, 1557
 glaucomatous optic neuropathy, 1538–1539
 visualization techniques on, 1523
- Optimal antimicrobial agent, 1335
- Optimal treatment, 1329
- OPTIMAM Mammography Image Database (OMI-DB), 1289
- Optimization-based approaches, 412
- Oral squamous cell carcinoma (OSCC), 911
- ORFéAD network, 1773
- Organ donor selection, 1390
- Orthognathic surgery, 899–900
- Orthopaedics, 874
 care, 874
 databases, 881–882
 features, 874
 fracture detection systems, 877
 implant detection, 878
 knee pathology detection and segmentation, 877
 musculoskeletal image acquisition, 876
 musculoskeletal image scheduling and protocolling, 876
 oncology detection, 878
 post-operative complications and rehabilitation, 882–883
 surgery, 222
- Oscillometric method, 691
- Osteoarthritis (OA), 777–779, 877
- Osteoporosis, 679–681, 774, 781

- Otology
 ABR, 992
 auditory brainstem response (ABR), 992
 balance/vestibular pathologies, 993
 cochlear implant (CI), 993
 evoked compound action potentials (ECAPs), 993
 hearing impairment technologies, 993
 imaging modalities/radiomics, 991, 992
 noise induced hearing loss (NIHL), 992
 sensorineural hearing loss (SNHL), 992
 SSH, 992
 sudden sensorineural hearing loss (SSNHL), 992
- Outcome prediction, 906, 912–913
- Out-of-office BP, 691, 695
- Outpatient setting, 1334
- Over-testing, 275
- Oxygen desaturation index (ODI), 765
- P**
- Pacemakers, 1480, 1481, 1483, 1484
- Pan-allelic model, 1391
- PANDA, 495
- Pandemics, 1392
- Papilledema and optic disc abnormalities, 1527, 1530
- Pap smear test
 automated segmentation, 1322
 class identification and quantification, 1320
 database generation, 1321
 feature extraction, 1324
 feature selection, 1324
 ground truth labeling, 1321
 HPV vaccination, 1318
 LBC method, 1319
 MSER algorithm, 1321
 multi-class classification, 1324
 for quantification, 1324
 SIL, 1320
- Parasitemia, 1355
- Parkinson's disease, 1676, 1693
 electroencephalographic diagnosis, 1683
 ethical and social implications, 1686
 in-vivo detection and management, 1676–1686
 medical management drug repurposing, 1683
 surgical management, 1685–1686
- Partial nephrectomy, 866
- Participatory approaches, 1337
- Part-of-speech (POS), 272
- Passive diagnosis, 184, 193
- Pathogens, 1370–1375
- Pathologist's role, 525
- Pathology, 730
- Pathomics, 1267
- Patient activation measurement (PAM), 1237
- Patient and Public Involvement and Engagement (PPIE), 358
- Patient centeredness, 1126
- Patient Centered Outcomes Research Institute (PCORI), 607
- Patient-driven systems, 707
- Patient engagement, 360
- Patient experience, 356
- Patients' blood management system, 809
- Patient safety
 ACSNQIP, 217
 AI techniques, 216
 database, 222
 EHR, 217
 global public health, 216
 intelligent systems, 218
 preventing errors, 217
 quality, 222
 TI, 216
 treatment, 221
- Patient's perspective and AIM
 benefits and risks, 356–357
 blockchain, 359–360
 dynamic consent, 360
 federated computing, 360
 need for transparency, 352–353
 privacy by design, 358–359
 public's trust, 358
 public's understanding of AI, 354–355
 public's view on data sharing, 355–356
 regulations and data sharing, 353–354
- Patient Virtual Health Assistance, 619
- Pattern(s), 374
- Pattern recognition, 77, 552, 1204, 1206, 1207
 algorithms, 184
- Pauli X gate, 433, 434
- Pedagogical efforts, 1337
- Pediatric(s)
 AI solutions, 1041
 AI tools, 1030
 bone age, 495
 cardiology, 1032
 challenges, 1031, 1032
 definition, 1030
 endocrinology, 1035
 gastroenterology, 1040
 genetics, 1035
 neonatology, 1036, 1037
 ophthalmology, 1038
 PICU, 1040
 primary care, 1039
 radiology, 1039
 respiratory, 1033, 1034
- Pediatric Early Warning Score (PEWS), 1052
- Pediatric intensive care units (PICU), 1040
 automated vital sign pattern analyses, 1052, 1053
 CDS, 1053
 clinical assessment, 1048
 early detection, Sepsis, 1052
 infections, 1048
 ML model, 1053
 neonatal sepsis, 1050, 1051
 sepsis, 1049
 vital signs, 1049, 1050

- Perceptrons, 16
- Percutaneous nephrolithotomy (PCNL), 867
- Performance metrics, 1400
- Performance of AI, 234–235
- Perfusion and ischemic core mismatch paradigm (PWI/DWI mismatch), 1739
- Perfusion-diffusion mismatch paradigm, 1742
- Perfusion imaging-based TIA detection algorithm, 1745
- Perfusion imaging technology, 1742
- Periprocedural support, 468
- Personalized medicine, 690, 694, 697, 728, 1770–1771
- Pharmacokinetics, 638–639
- Pharmacokinetics and pharmacodynamics (PK/PD), 1328
- Pharmacologic management, 1445
- Pharmacology, 1255
- ADME in pharmacokinetics, 638–639
 - and chemical genomics-based approach, 638
 - and ligand screening, 636–637
 - omics data, 640
 - real-world data, 640–641
- Pharmacovigilance, 1346
- Phenotypic screening, 1392
- Philosophy, 372
- PHOG, 1794
- Physera, 1794
- Physical rehabilitation, 1791
- Physical therapy, 1811, 1815
- Physician Clinical Decision Support, 619
- Physician-patient relationship, 142–143
- Physicians, 1221, 1223, 1226
- Physiological based pharmacokinetic (PBPK) modelling, 1498–1499
- Physiologically based models, 582
- Physiological measurements, 1470
- Physiotherapy, AI for, 727
- activity of daily living monitoring, 1800
 - cognitive impaired patients, 1803
 - community physiotherapy and care, 1803
 - companies involved in, 1795–1798
 - education and use of simulation, 1799–1800
 - exergames and serious games, 1793–1799
 - functional and feedback systems, 1803
 - future, 1804
 - hemi-neglect training, 1802
 - inattention monitoring, 1802
 - physical patient training, 1791–1793
 - respiratory management, 1802–1803
 - robotic physiotherapy, 1800
 - smart watches and wearables, 1804
 - virtual reality platforms, 1801
 - wheelchair users and assisted mobility support, 1801
- Physitrack, 1794
- PICCOLO, 975
- Picture archiving and communication system (PACS), 466
- Pinocytic-clathrin-mediated transport, 1178
- Pittsburgh urban mathematics project (PUMP), 333
- Pituitary and adrenal disorders, AI applications
- diagnosis and subtyping, 682
 - prediction of treatment outcomes, 682
- Pixel-based or feature-based alignment methods, 945
- Pixel-level thresholding, 530
- Plan evaluation, 1280
- Planning, 906, 912
- Plasmodium*, 1332, 1354
- differential diagnosis, 1354
 - life stage detection, 1360
 - P. falciparum* detection, 1359–1360
- Plethysmography (PPG), 691
- 4P medicine, 595, 1348
- 10-point Houston Intra-Arterial Therapy (HIAT) score, 1512
- Point of care diagnostic tests, 1436
- Point-of-care ultrasound (PoCUS), 763
- Polarity, 1347
- Police custody, 1768
- Policy makers, 1350
- Pollutant dispersion, 628
- Polycythaemia rubra vera screening, 1430
- Polygenic risk scores (PRS), 1091, 1414
- Polyp-Alert system, 923
- Polyp classification, 969
- Polyp detection, 968, 973
- Polypharmacy, 221
- Polyp localization, 968
- Polyp segmentation, 968
- Polysomnography (PSG), 765
- Pooled Resource Open-Access ALS Clinical Trials, 1692
- Pooling layer, 79
- Population epidemiology, 1343
- Population health, 595, 1350
- Population Intervention, Comparison, Outcomes (PICO), 257
- Population level interventions, 121–122
- Population pharmacokinetic analysis (popPK), 1490
- PoseNet, 1803
- Positron emission tomography (PET), 503, 1265, 1585
- Post-hoc explanations, 508
- Post-hoc interpretability, 507
- Post-mortem identification, 1780–1781
- Post-mortem interval estimation, 1781
- Post-partum hemorrhage, 1435
- Post-traumatic stress disorder (PTSD)
- after exposure to traumatic events, 1630–1631
 - AI, characterization and diagnosis of, 1635
 - AI for the diagnosis and differentiation of, 1636–1637
- DSM-5 diagnostic criteria for, 1632
- early prediction of, 1633–1635
- evolution of, 1632
- genomic data, 1636
- neuro-imaging data, 1636, 1637
- prediction of response to treatment, 1635
- predictive factors and determinants of, 1630
- screening and diagnosis of, 1632–1633
- subtypes of, use of AI to characterise, 1637
- trends and challenges, 1638–1639

- Precision, 595
medicine, 290–291, 344–345, 713, 728, 732, 1570, 1572, 1575, 1576, 1604
psychiatry, 1604
public health, 594, 596, 599
- Prediction model Risk Of Bias Assessment Tool (PROBAST), 393
- Predictive analytics, 881–883, 1004, 1264
- Predictive data analytics, 375
- Predictive medicine, 690
- Predictors, 1631
- Preimplantation genetic diagnosis (PGD), 1022
- Pre-operative planning, 828
- Prevention of errors, 467–468
- Primary care, 1631, 1636
- Primary healthcare, 619, 714, 737
altered roles, 717
cardiac, 735
chronic neurological and neuropsychiatric disease monitoring, 736
electronic health records and data ownership, 715–716
endocrinology and diabetics, 735
gender aspects, 734
global companies, 718–721
global macrotrends, 716
hypertension, 735
legal and regulatory aspects, 732
machine learning algorithms, 722–731
medical imaging diagnostics and radiology, 732–733
medical informatics and clinical decision support, 733
obstetrics, pregnancy and pediatrics, 736
opportunities, 714
patient's perspective, 733–734
patient safety, 732
point of care dermatology and ophthalmology, 734
precision medicine, 732
privacy concerns, 732
public health aspects, 734
respiratory, 735–736
shift of balance, 714–715
symptom checkers and dissemination of specialities, 716–717
- Primary immunodeficiency diseases (PID), 1398, 1401
- Prime number factorization function, 430
- Principal component analysis (PCA), 82, 1206, 1207, 1474, 1494, 1496, 1668, 1757
- Privacy, 232
by design, 358–359
- Privacy-preserving machine learning (PPML), 149, 154
- Probabilistic causal discovery in sequential dataset (PCDSD) model, 1695, 1697, 1699
- Probabilistic clustering (PC), 1474
- Probabilistic graphical model, 274
- Probability Decision Score (*PDS*), 796
- Probe-based confocal laser endomicroscopy (pCLE), 845, 958
- Problem Areas in Diabetes (PAID), 1240
- Problem-solving, 364
- Prognostication, in Parkinson's Disease, 1683
- Progressive NAS search strategy, 418
- Progressive neural networks, 413
- Prostate cancer, 864–866
- Prostate cANcer graDe Assessment (PANDA), 529
- Prostate Imaging-Reporting and Data System (PIRADS), 865
- Protected Health Information (PHI), 272
- Protein(s), 1392
folding, 662–663
structure, 1390
structure prediction method, 663
- Proteomic(s), 1062
analysis of vectors, 1365
- Prototype feature vector, 412
- Proximal-interphalangeal (PIP) joints, 780
- Pseudomonas aeruginosa*, 1332
- Pseudonymization, 148, 493
- Psoriasis, 554–556
- Psoriasis Area and Severity Index (PASI), 554
- Psychiatric disorder, 1631
- Psychosis, 30, 1597, 1598, 1601, 1603, 1604
- Psychosomatic therapy, 1254
- Psychotic disorders, 1596, 1597, 1602, 1604
- Public health, 594, 595
AI, 597, 600
data, 596
future challenges, 600
individual/social preference, 601
LHS, 597
patient, 597
precision, 595
private sector, 599
research/governance, 599
risk/insurance, 598
segmentation/targeting, 598
surveillance, 1345
surveillance systems, 597
- Public trust, 352, 354, 357, 358
- Pulmonary function tests (PFTs), 763, 766
ML, 764
spirometry, 764
- Pulse rate variability, 1149
- Pulse transit time (PTT), 691
- Pyruvate kinase deficiency screening, 1430
- Python, 7–8
- Q**
- Quality adjusted life years (QALYs), 1031
- Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2), 392
- Quality assessment tools, AI, 387
PROBAST-AI, 393
QUADAS-AI, 392–393
- Quality assurance (QA), 1276, 1280
- Quality of life (QoL) assessment, 581, 1126, 1128
- Quality score, 1133
- Auquantitative reverse transcriptase PCR (qRT-PCR), 1113

- Quantitative structure-activity relationship-like (QSAR)
approach, 637, 1492
- Quantum annealing, 441
- Quantum biology, 441
- Quantum computer types, 427
- Quantum computing
algorithms, 438–441
basics of, 426–429
Bernstein-Vazirani algorithm, 440
vs. classical computing circuit, 429–433
companies developing, 429
future translational aspects for e-healthcare, 442
in healthcare, 441
history of, 426
models, 428
quantum tunnelling, 435–436
reinforcement learning, 441
Shor’s algorithm, 439
- Quantum entanglement, 436
- Quantum gates, 433, 434
- Quantum hidden Markov chain algorithms, 440
- Quantum natural language processing algorithms, 440
- Quantum neurosurgery, 1182
- Quantum superposition, 433
- Quantum tunnelling, 435–436
- Qubits, 431
- Quick Sequential Organ Failure Assessment (qSOFA), 1474
- QuikSCAT, 630
- R**
- Radial Basis Function (RBF), 583, 586
- Radiation dose delivery, 1276–1278, 1280, 1282, 1283, 1285
- Radiation oncology, 1269, 1278
- Radiation treatment planning, 1278
- Radiohistomics, 1268
- Radiology, 525
- Radiomics, 1265
- Radiotherapy
aim, 1276
cancer patients, 1284
chain, 1277, 1285
definition, 1276
fluence map, 1282
intensity modulated, 1276, 1282
patient care, 1284
photon beam segment, 1282
process, 1280
- Raman optical spectroscopy, 1331
- Raman spectroscopy (RS), 846, 1332
- Random forest, 13, 248, 586, 1115, 1206, 1389, 1474, 1495
algorithm, 626
analysis, 1332
Model, 1332
- Random forest (RF)-based methods, 693
- Randomised controlled trials, 1461
- Randomised controlled trials (RCT), 1235
- Randomized controlled trials (RCTs), 647, 1235, 1461
- Randomized controlled clinical trial, 1446
- Randomized-control methodology, 1337
- Random survival forests (RSF), 779
- RAPIDAI, 1225
- Rapidly growing markets, 66
- RAPID software, 1741
- Ratio of means meta-analysis, 331
- 3rd Generation (3G) mobile networks, 1231
- Realization, 1337
- Real-world data, 263, 640–641
- Real-world evidence, 263
- Real-world validation, 1223
- Reasoning, 271–275, 280, 284
- Receiver operating characteristic (ROC) analysis, 1295
- Receiver operating characteristic (ROC) curve, 572
- Recognition, 378
- Rectifying linear unit ‘ filter (ReLU), 79
- Recurrent neural network (RNN), 18, 43–44, 247, 249, 455, 585, 626, 760, 775, 818, 1392, 1586
approach, 1380
- Recursive neural knowledge network (RNKN), 281
- Red Blood Cell (RBC), 581
- Red green blue (RGB) values, 530
- Regenerative medicine, 1181
- Regional-based Convolutional Neural Networks (R-CNN), 956
- Region-based CNN (R-CNN), 542, 545
- Region-based techniques, 1721
- Region of interest (ROI), 461, 761
- Regression problems, 11
- Regulation, 352, 355, 357
- Rehabilitation, 1810
AI wearable monitoring devices, 1815
assessment and decision support systems, AI in, 1814
benefits of, 1810
definition, 1810
gait rehabilitation, 1813–1814
LfD, 1813
neuropsychology, 1811–1812
occupational rehabilitation, 1812
phases of, 1810–1811
physical therapy, 1811
rehabilitation prognosis, AI in, 1814–1815
speech therapy, 1811
virtual reality and serious games, AI in, 1815–1816
- Rehabilitation routine mining
evaluation, 312–314
methodology, 312
- Re-identification, 353, 359
- Reinforcement learning (RL), 15–16, 196, 410, 441, 587, 891, 1058, 1100, 1374, 1446–1447, 1664
algorithms, 477
for sequential decision making, 168–169
- Reinforcement machine learning (RML), 749–751
- Relapse, 1574
- Relational practice, 379
- ReLeaSE, 1062

- Relevance vector machines (RVM), 1474
Remote health monitoring, 1152
Remote photoplethysmography, 1147
Remote-sensing satellite sensors, 629, 631
Renal pathology, 544
Reperfusion therapy, 1737
Reporting metrics, 973
Reporting standards, AI, 386
 SPIRIT-AI and CONSORT-AI, 389
 STARD-AI, 389–392
 TRIPOD-AI, 392
Reproductive medicine, AI inopportunities and limitations, *see* Artificial intelligence (AI)
Reproductive medicine (RM), 1018
 AI relationship, 1020, 1021
 CRISPR, 1021
 genetic editing technology, 1023
 history of, 1018–1020
 male experience, 1023
 PGD, 1022
 transgender and gender nonbinary inclusion, 1024
 transgender, gender fluid and non-binary medicine, 1024
Residency and fellowship training, 469–470
ResNet, 1308
Respect for autonomy, 139–140
Respiratory medicine, 760, 762, 765, 766, 1237–1238
Respiratory Syncytial Virus (RSV), 1333
Response Assessment in Neuro-Oncology (RANO), 1726, 1727
Restricted Boltzmann machines (RBMs), 248
Restricted isometry property (RIP) conditions, 47
Retcam imaging, 1530
Retinal fundus photographs
 age-related macular degeneration, 1529
 diabetic retinopathy, 1521–1528
 glaucoma, 1528–1529
 papilledema and optic disc abnormalities, 1530
 retinopathy of prematurity, 1529
 systemic diseases, 1530–1534
Retinopathy of prematurity (ROP), 211, 1038, 1527, 1529
Reverse time attention model (RETAINT), 249
Reynolds-averaged Navier–Stokes simulation (RANS), 625
Rheumatoid arthritis (RA), 774, 777, 1238, 1405
Rheumatology, artificial intelligence, 776, 777, 1238–1239
 EHRs, 776–779
 genetic and biomarker data, 779–780
 medical imaging, 780–781
 mixture data, 781–782
Rhinology
 Chronic rhinosinusitis (CRS), 994
 CRS, 994
 definition, 994
 imaging, 994
 pathological diagnosis, 994
 Sino-Nasal Outcome Test-22 (SNOT-22), 994
Ripplet transform analysis, 1324
Risk factor analysis, 595, 598
Risk prediction, 275, 727, 858–859
Risk stratification, 707
R language, 8
RNA sequencing (RNA-seq) protocols, 1112, 1392
Robot assisted radical prostatectomies (RARP), 335
Robotics, 123–124, 745, 753, 755, 1810
 autonomy, 878
 continued adoption of, 881
 physiotherapy, 1800
 in rehabilitation, 1812–1814
 surgery, 826
 technology, 754, 755
Robotic-assisted AI, 1680
Robotic-assisted laparoscopic radical prostatectomy, 864
Root mean square error (RMSE), 627, 692, 1281
Rudimentary machine, 1264
Rule-based decision support system, 68
Rule-based diagnosis, 186
Rule-based reasoning (RBR), 702
Rule-based systems, 587
Rules based model (RBM), 1333
Rural health, 606
Rwanda, 618
- S**
- Saliency maps, 464
SARS-CoV2, 1374
SARS-CoV-2, 1390
Satellite sensors, 630
 active sensor, 630
 passive sensors, 630
Satisfaction of search (SoS) bias, 184
Savitzky-Golay filtering, 1211
Scalability, 413
Scale invariant feature transforms (SIFT), 1793
Scanning tunnelling microscopy, 436
Schizophrenia, 23, 1596–1605
 ANNs, 1599
 non-verbal social interactions, 1600
 SVMs, 1598, 1601, 1602
 symbolic and subsymbolic artificial intelligence, 1598
Screening, 259
 automation, 260
Search strategies, 259
SEASAT satellite, 630
Seattle Heart Failure Model (SHFM), 817
Secondary structure prediction, 663
Second generation (2G) mobile networks, 1231
Second order cone (SOCP) programming problems, 794
Secure Multi-Party Computation (SMPC), 153
Sedatives, 1473
Segmentation, 866, 909, 923, 925, 931
SE Health, 1616
Seizure monitoring systems, 736
Seizure(s), 1755
 monitoring systems, 736
 See also Epilepsy

- Select and test model, 183
 Self-organizing map (SOM) models, 590, 1268, 1496
 Self-supervised methods, 839
 Semantic network, 1597
 Semen analyses (SA), 869
 Semi-automation, 124
 Semi-autonomous intraoperative robotics, 879
 Semi-supervised learning, 1058, 1722, 1723
 Sensor drift, 1206, 1210–1212
 Sentiment analysis, 1347
 Sepsis, 1333, 1334, 1374, 1472–1474
 Sequence-based models, 1392
 Sequence mapping, 245
 Sequential minimal optimization (SMO), 1212
 Sequential [Sepsis-related] Organ Failure Assessment (SOFA), 1049
 Serum creatinine (SCr), 562, 563, 807
 Severity recognition, 1581
 Sex and gender bias
 in machine learning models, 399–400
 in medicine, 398–399
 role in machine learning models for medicine, 400–404
 Sexually transmitted infections (STI), 1318
 Shape-from-shading (SfS), 945
 SHapley Additive exPlanations (SHAP), 249, 507, 573, 694, 931
 Shared decision-making, 256, 264, 265
 Shor's algorithm, 439
Silhouette score, 1133
 Silico tools, 1390
 Silico vaccine discovery, 1389
 Simplified Molecular Input Line Entry System (SMILES), 1375
 Simulated learning environments (SLE), 1799
 Simultaneous localization and mapping (SLAM), 841, 945
 Singapore I vessel assessment-DL system (SIVA-DLS), 1531
 Single-cell cancerous subclones, 1432
 Single nucleoside polymorphism (SNP), 1091, 1112, 1404, 1484
 Single Shot MultiBox Detector (SSD) deep learning model, 1104
 Singular value decomposition (SVD), 628, 841, 1211
 SIR-type models, 1380
 Skewness, 788
 Skin cancer, 552–554, 1225
 Skin lesion classification model, 401
 Sleep monitoring, 765
 Sluice networks, 415
 Smallpox, 1388
 SmartAQnet, 633
 Smartphone app, 1230, 1233
 Smartphones, 692, 1146, 1153
 Smart textiles, 1196
 Social analytics for healthcare, 1127
 Social dilemma, 131
 Socially interactive robots, 1600
 Social media analysis, 1347
 Social medicine, 594, 595
 Social phenotype of diseases, 1126
 Societal change, 753
 Society of American Gastrointestinal and Endoscopic Surgeons (SAGES), 860
 Society of Critical Care Medicine (SCCM), 1049
 Soft parameter sharing, 414
 Software as a medical device (SaMD), 932
 Software for Primary Immunodeficiency Recognition Intervention and Tracking (SPIRIT), 1401
 Software guard extensions (SGX), 154
 Somatosensory evoked potentials, 1762
 Sorenson-Dice coefficient, 866
 Sparse analysis techniques, 794
 BBLL-R, 795
 BBLL-S, 796
 BBMAP, 795
 block decomposition, 794
 ensemble classification, 795
 Sparse modeling methods, 792
 Sparse representation classification (SRC) method, 793
 Sparse representation problem, 794
 SPARSity based super-resolution Correlation Microscopy (SPARCOM), 49
 Sparsity concentration index (SCI), 794
 Speech therapy, 1811
 Spinal deformity, 882
 Spinal stratification, 1692
 Spin-Q technology, 429
 SPIRIT-AI, 532
 Spirometry, 760, 763, 764, 766
 Spleen deficiency syndrome, 1256
 Split learning, 1161
 Splus statistical programming language approach, 1103
 Sporadic stratification, 1692
 Sports hematology, 1437
 Squamous cell carcinoma (SCC), 526, 848, 943
 Squamous intra-epithelial lesion (SIL), 1319
 Stability-Plasticity Dilemma, 413
 Standard deviation (SD), 692
 Standardized data, 1336
 Standardized mean difference, 327
 Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT), 389, 657
 Standards for Reporting of Diagnostic Accuracy Studies (STARD), 389–392
 STARD-AI, 532
 Statistical approach, 1349
 Statistical methods, 1769
 Statistical significance, 1335
 Stem cells, 1098
 AI approaches in prediction, validation and analysis of, 1099
 AI implementation in progression of, 1100
 applications, 1099
 types, 1098
 Stemming, 19
 Stochastic Gradient Descent (SGD), 584
 Strategic Anemia Advisor (SAA), 583
 Stress assessment, 1146

- Stroke, 1224, 1504, 1734
 acute stroke therapy, 1737–1743
 care workflow, 1735
 early symptom detection, 1736–1738
- Strong AI, 7
- Structural causal models (SCMs), 172
- Structured electronic medical record (SEMUR), 1250
- Structure-from-motion (SfM), 838, 840, 945
- Structure learning, 1349
- Substance dependence, 1620, 1623, 1625
- Substance use disorder ((SUD), 1620
 challenges in diagnosis and treatment, 1620–1622
 machine learning, 1624
 role of artificial intelligence, 1622–1626
- Subsymbolic artificial intelligence, 1598
- Sudden cardiac arrest (SCA), 1480–1482
- Sudden unexpected death in epilepsy (SUDEP), 1238
- Suicidality, 1574
- SUMEX-AIM project, 1060
- Super AI, 7
- SuperCooperators, 132, 133
- Super learner algorithm, 1745
- Supervised and unsupervised learning, 1757
- Supervised learning, 11–15, 165, 409, 838, 857, 881, 891, 1100, 1114, 1493–1494, 1568, 1570
- Supervised machine learning (SML), 748–749, 1332, 1334, 1507
- Supervised meta learning process, 410
- Supervised models, 1664
- Supervised/semi-supervised/unsupervised learning modes, 1266
- Supervised transformation estimation, 829
- Support dilution, 304, 306
- Support vector machines (SVM), 14–15, 23, 166, 194, 195, 530, 541, 545, 546, 589, 847, 943, 1011, 1088, 1117, 1206, 1281, 1389, 1401, 1445, 1474, 1495, 1598–1602
 classification, 953, 1332
- Surgeon’s skill, 831
- Surgery, artificial intelligence
 computer vision, 857–858
 intraoperative video analysis, 859–860
 NLP, 858
 preoperative risk prediction, 858–859
 regulatory and legal considerations, 860–861
 surgical workflow analysis, 860
- Surgical automation, 857
- Surgical Control Tower (SCT), 903
- Surgical Data Science (SDS), 830, 942
- Surgical interventions, 123
- Surgical pathology
 detecting and classifying disease, 525, 526
 finding/outlining tumorous and tissue, 527
 finding and highlighting small objects, 528
 grading and scoring disease, 525
 predictive tasks, 528, 529
- Surgical planning, 826, 827
- Surgical robotics, 826
 AR, 844
 cognitive, 837
 context aware decision support, 846–848
 depth perception, 837–840
 external manipulations, 836
 haptic feedback/tissue interaction sensing, 843, 844
 Robot-assisted tissue scanning, 846
 safety, 849
 SfM, 840, 841
 SLAM, 841, 842
 tool tracking, 843
- Surgical-skill assessment, 831
- Surrogate models, 691
- Surveillance of infection, 1336
- Surveillance systems, 1336
- Sustainable Development Goals (SDG), 615
- SWI Prolog*, 8
- SWORD, 1794
- Symbolic artificial intelligence, 1603
- Symbolic-connectionism, 268–272
- Symbolic languages
 calculator, 205
 computer, 206
 Hurufism, 205
- Symbolic reasoning algorithms, 1597
- Sympathetic and parasympathetic nervous system activity, 1049
- Sympathetic nervous system activity, 1151
- Syndrome, 1251
 biomechanisms, 1256
 TCM, 1253
- Systematic biases, 1337
- Systematic reviews (SR), 257, 258
 aim, 261
 automation, 258
 living, 260, 265
 methodologies, 260
- System-based approach, 217
- Systemic diseases, 1530–1534
- Systemic lupus erythematosus (SLE), 774, 779, 780
- Systemic sclerosis, 774, 782
- Systems biology, 1101
- System vaccinology, 1389
- Systolic Blood Pressure Intervention Trial (SPRINT), 693
- Systolic BP, 691
- T**
- T1D, 702, 703, 705
- T2D, 702, 703, 705
- Tachycardia, 818
- Targeting aging with metformin (TAME) trial, 1160
- Target task, 973
- Task automation, 727
- Task bias, 401
- Taxonomy alignment, 1079
 merge algorithm, 1079
 pruning algorithm, 1080
 split algorithm, 1080

- t-distributed stochastic neighbor embedding (t-SNE)
technique, 281
- Technological advances, 702
- Teleconsultation, 1221
- Teledermatology (TD), 557, 1225
- Telediagnosis, 1221
- Telehealth, 1221, 1230
- Telemedicine, 61, 557, 725, 760, 764, 1153, 1220
AI, 1222
definition, 1221
education, 1226
limitations, 1221
modalities, 1221
telehealth, 1221
teleoftalmology, 1224
training, 1226
- Telemonitoring system, 765, 1221
- Teleoftalmology, 1224
- Teleoncology, 1222
- TELEOS project, 334
- Telepathology, 523, 524
- Telestroke, 1225, 1512, 1513
- Teletriage, 1221
- Template matching, 1510
- Temporal lobe epilepsy, 1761
- Temporal Pyramid Network (TPN), 1588
- Temporomandibular joint (TMJ) surgery, 900–901
- Tess, 1611
- Test set technologies, COVID-19, 512–513
- Textual data, 1347
- Texture-based classification, 786, 797
feature computation, 786–790–791
- Texture-based classifiers, 798
- Thalassaemia, 1429
- Thanatology, 1783
- The end of the theory, 1349
- The Origin of Species*, 344
- Theory of iatro-mathematics, 5
- The Population Research in Identities and Disparities for Equality* (PRIDE), 1025
- Thermalytix, 1312
- Thermalytix risk estimation (TRS), 1311
- Threat model, 147
- Thrombolysis, 1739
- Thrombolysis in Cerebral Infarction (TICI)
score, 1512
- Thrombolytic agents, 1504
- Thucydides, 1388
- Thyroid and endocrine surgery
clinicopathological prediction, 990
cytopathological diagnosis, 990
hyperparathyroidism, 991
MRI radiomics, 990
thyroid cancer, 990
ultrasound, 990
- Thyroid cancer, AI application, 681
- Time-lapse imaging (TLI) systems, 1012
- Time-series data, 571, 692, 697
- Tokenization, 19
- Tooth, 907, 909, 911, 912
- Topic modeling, 1127, 1130
agglomerative hierarchical clustering, 1131–1132
dimensionality reduction, 1130–1131
summarization, 1132
- Tort law, 134
automated systems, 134
computer generated torts, 135
3D printing, 136
legal liability, 136
negligence standard, 135
patient safety standards, 136
smart regulation, 136
social dilemmas, 137
strict liability, 135
- Total hip arthroplasty, 881
- Totalled Health Risks in Vascular Events (THRIVE)
score, 1512
- Toxicovigilance, 1489–1490
- TPF measures, 1295
- Traditional approaches, 1392
- Traditional Chinese medicine language system
(TCMLS), 1250
- Traditional Chinese Medicine (TCM), 1248, 1249
- Traditional cohorts, 1347
- Training, 1569, 1573
- Transcatheter aortic valve implantation (TAVI), 818
- Transcutaneous bilirubin (TcB), 1037
- Transdermal optical imaging, 729, 1146–1147
accurate blood pressure measurements, 1151–1152
biomechanics and video-capture, 1147
challenges for healthcare delivery, 1152–1154
future uses and challenges of, 1154–1155
health care quality and efficiency, 1152
heart rate and heart rate variability, 1151–1152
photoplethysmographic signal extraction, 1148
pulse rate, 1149
pulse rate variability, 1149
scientific foundations of, 1147
systolic and diastolic blood pressures, 1149
telemedicine, 1153
tools, 1153–1154
trends and potential impact of, 1152–1154
- Transfer learning techniques, 410, 1359, 1723
- Transfusion medicine, 1435
- Transient ischemic attack (TIA) management, 1745
- Transparency, 228, 229, 236, 352, 354, 356–358
- Transparent models, 506
- Transparent Reporting of a multivariable prediction
model for Individual Prognosis Or Diagnosis
(TRIPOD), 392, 1008
- Transplantation, 1099–1101, 1104, 1389, 1390
- Transplant rejection, 1390
- Trauma, 874
surgery, 901–902
- Traumatic events, 1630–1631
- Treatment delay, 1571

- Treatment injury (TI), 216, 223
Treatment pathways, 125
Treatment planning system, 1276, 1278
TREAT system, 1335
Triage system, 1505, 1513
Trial of Org 10172 in Acute Stroke Treatment (TOAST) criteria, 1743
TRIPOD-AI, 532
Trophectoderm biopsy, 1012
trRosetta, 1390
Trusted Execution Environments (TEEs), 154
Trustworthiness, 506, 508
Tuberculosis (TB), 1334–1336
Tubules, 540, 542, 543
Tumour marker testing, 809
Tumour mutational burden (TMB), 1431
 β -turns, 663
Twitter, 1347
Two-module transfer learning weighted optimised deformable convolutional neural networks (TWO-DCNN), 1429
Type 1 reasoning, 183
Type 2 diabetes mellitus (T2DM), 1240
Type 2 reasoning, 183
- U**
Ultrasonography (US), 1265
Ultrasound, 1303–1304
U-Net architecture, 1509
U-Net-based algorithm, 1104
Unifocal-onset epilepsy, 1755
Unipolar depressive disorder, 1572
United States Preventive Services Task Force (USPSTF), 264
Unsupervised and supervised learning methods, 1433
Unsupervised clustering algorithms, 1417
Unsupervised learning, 166, 857–858, 881, 891, 1058, 1100, 1115, 1494–1496, 1570, 1664, 1722, 1723, 1757 algorithm, 15
Unsupervised machine learning (UML), 749, 1334, 1507
Unsupervised transformation estimation, 829
Urban-Heat-Island (UHI) effect, 629
Urologic oncology history, 864
kidney cancer, 866
prostate cancer, 864–866
urothelial cancer, 866–867
Urology, artificial intelligence, 1239 andrology, 868–870
endourology, 867–868
urologic oncology, 864–867
Uropathology, 525
Urothelial cancer, 866–867
US Food and drug administration (FDA), 533
Utilization strategies of change, 1612–1613
- V**
Vaccination, 345 campaigns, 1383
Vaccine discovery challenges, 1390 distribution and planning, 1389 search steps, 1389
Vaccinia virus, 1373
Validation, 387
Value-based care models, 607
Value metrics, 607
Value of information (VOI), 170, 191
Vanishing gradient problem, 530
Variant call format (VCF), 1392
Variational autoencoders, 668
Variational autoencoder (VAE), 1089
VARK model, 322
Vascular deformations, 1310
Vascular dementia, 1681
Vascular interventions, 465
Vasodilators infusion, 1458
Vasopressors titration, 1458
Vector control, 1355, 1366
Ventilator-induced lung injury (VILI), 1473
Ventilatory-associated lung injury (VALI), 1473
Ventricular arrhythmia, 1480–1481
Ventricular fibrillation, 1480, 1481
Ventricular tachycardia, 818, 1480–1483
Verbal Intelligence Quotient, 1582
Vertical root fractures (VRFs), 911
Vertigo, 1706, 1707
VGGNet, 50
Video-based AI rating, 1683
Video capsule endoscopy (VCE), 921
Violence, 1631, 1635, 1768
Virology laboratory, 1332
Virtual-fixture control, 826
Virtual microscopy, 523
Virtual reality, 946, 1801, 1815
Virtual screening, 1061
Viruses, 1370–1375
Visual field (VF), 211, 1539–1540
Visualization tools, 1337
Volatile organic compounds (VOCs), 765, 1210
Volatiles, 1204, 1206, 1210
Volatilome, 1206, 1210
Volocity® cytoplasmic cell marker, 1103
Volumetric laser endomicroscopy (VLE), 959
Volumetric modulated arc therapy, 1282, 1283
Voxel-based dose prediction algorithm, 1279
- W**
WannaCry attack, 299
Warfarin, 815
Wastewater data, 1345

- Watson, 5, 7
 for genomics, 1436
 project, 64
- Wavelet texture descriptor, 787
- Weak AI, 7
- Wearable(s), 120, 121, 1484
 computing, 1193
 devices, 692, 697
- Wearable medical devices
 AI-assisted fabrication of wearables, 1196
 design and construction, 1194–1197
 personalization, 1195–1196
 prefabrication and process planning, 1196
 safety assessment, 1197
 usability, 1199
 user acceptance, 1199
 validation, 1197–1199
- Wearable monitoring devices, 1815
 fall detection, 1815
 monitoring purposes, 1815
- Web-based system, 1285
- Web search engine, 258
- Weibull model, 808
- Weighted mean difference, 327
- Wellness, 1611
- White light endoscopy (WLE), 952
- Whole-genome sequencing (WGS), 1331
- Whole slide images (WSIs), 525, 530, 541
- Whole slide imaging, 523, 524, 530
- Wireless capsule endoscopy (WCE), 943
- Working channel, 940
- Working length (WL), 910
- World Confederation for Physical Therapy (WCPT), 1811
- World Health Organization (WHO), 219, 1025
- Wrapper approach, 245
- X**
- XenoSite, 639
- XGBoost, 696
- X-learner, 694
- X-ray images, 760
- X-ray imaging and endoscopy, 829
- Z**
- Zuri, 1615